Construction automatique d'un dictionnaire des événements d'une vidéo

par

Ouael Chaari

Mémoire présenté au Département d'informatique
en vue de l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, janvier 2010

Your file Votre référence
ISBN: 978-0-494-61460-0
Our file Notre référence
ISBN: 978-0-494-61460-0

Le 18 janvier 2010

le jury a accepté le mémoire de Monsieur Ouael Chaari
dans sa version finale.

Membres du jury

Professeur Djemel Ziou
Directeur de recherche
Département d'informatique

Professeur Mohand Said Allili
Membre
Département d'informatique

Professeur Jean-Pierre Dussault
Président rapporteur
Département d'informatique

SOMMAIRE

L'interprétation automatique d'événements des objets dans la vidéo est un sujet de recherche en croissance. Les humains cherchent de plus en plus à doter les systèmes d'informatique d'une intelligence pour la prise de décisions. Malgré l'engouement pour cette recherche, il reste encore plusieurs problèmes et défis à relever surtout en ce qui concerne la généralisation, la précision et l'automatisation d'un système de reconnaissance d'événements dans la vidéo. Nous proposons un système SIFD automatique de la vidéo dont l'objectif est l'obtention d'un dictionnaire de la vidéo qui décrit dans un langage naturel le contenu de la vidéo. Dans ce système, un modèle de reconnaissance d'actions humaines est développé. Les résultats obtenus pour ce modèle montrent sa robustesse et ses excellentes performances par rapport aux autres travaux existants. La contribution de notre travail réside dans une nouvelle caractéristique CSST et dans la formation du dictionnaire.

REMERCIEMENTS

J'aimerais remercier Mr Djemel Ziou, mon directeur de maitrise pour l'aide qu'il m'a apportée tout au long de mon cheminement de maitrise. Ses précieux conseils et son soutien m'ont été d'un grand profit.

Je remercie aussi mon ami Omar pour sa collaboration dans ce travail et pour l'échange qu'on a eu. Un grand merci à mes collègues au centre de recherche MOIVRE, surtout Hamza, Riadh et Sabri, pour leurs aides précieuses et leurs présences.

Je remercie du fond de mon coeur, mes parents qui sans leur soutien, je n'aurais jamais eu cette chance de faire ma maitrise. Un grand merci à mon frère, à ma soeur et spécialement à Btissam, pour leur présence quotidienne.

Un grand merci à mes ami(e)s et à tous ceux qui m'ont aidé de proche ou de loin pour m'avoir supporté et soutenu durant mes études.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

Le traitement de vidéos désigne l'étude et la transformation de vidéos numériques afin d'améliorer leur qualité, de réduire leur coût de stockage ou d'en extraire les informations pertinentes. Le traitement de vidéos a été développé surtout pour répondre à des problèmes reliés à la structure (plans, scènes, etc.) et à l'enchainement des images (mouvement, suivi, etc.). Depuis les années 1990, la vidéo est utilisée dans l'industrie, le cinéma, la télévision, la robotique, la sécurité, le web pour ne citer que ceux-là.

Les algorithmes de traitement de vidéos se confondent se divisent en trois groupes : le bas niveau, le niveau structurel et le haut niveau [41]. Le bas niveau correspond au traitement de chaque image de la vidéo et donc de chacun des pixels (extraire les objets, les zones d'intérêts, le mouvement, etc.). Le niveau intermédiaire s'occupe du traitement d'un ensemble d'images et par conséquent d'un ensemble de pixels (extraire les plans, les scènes, etc.). Le haut niveau est celui qui fournit une information sémantique sur la vidéo. L'application de ce niveau implique l'utilisation d'autres algorithmes de bas niveau et de niveau intermédiaire (extraction des caractéristiques, extraction des plans, etc.).

Dans notre mémoire, nous nous intéressons principalement à la formation d'un dictionnaire des événements de la vidéo. Le dictionnaire décrira le contenu visuel de la vidéo en terme d'événements et de leur sémantique. Les utilisateurs du dictionnaire sont nombreux : l'archivage des documents visuels, le résumé des films, le résumé des vidéosur-

veillances (reconnaissance d'actions), etc. Dans ce mémoire nous nous intéressons à la construction d'un dictionnaire des actions humaines. Nous cherchons à décrire les actions humaines qui se trouvent dans une vidéo et cela en distinguant les mouvements et leurs trajectoires.

Les principales contributions de notre travail se récapitulent en deux points. Premièrement, nous proposons une nouvelle méthode pour la reconnaissance d'actions humaines fondée sur l'extraction et l'utilisation de nouvelles caractéristiques et de modèles statistiques pour la classification. Deuxièmement, nous avons implanté et validé un système d'interprétation d'événements pour la fabrication du dictionnaire.

Les caractéristiques sont extraites des images de la vidéo et peuvent être d'origine spatiale ou temporelle. Nous avons proposé une nouvelle caractéristique la CSST. Cette caractéristique englobe une information spatio-temporelle qui représente une région de la vidéo appelée zone d'intérêt. Notons que l'extraction de caractéristiques et la classification ont été realisées conjointement avec Omar Chahid.

L'interprétation des événements de la vidéo se base sur l'extraction, l'interprétation et la mise en relation des actions. Ainsi, l'extraction des actions se fait selon des méthodes de suivi et de détection de mouvement d'objets. Alors que l'interprétation se fait par l'intermédiaire de modèles de classification (KNN, K-means, etc.) ou par des méthodes d'étiquetages manuelles (intervention de l'utilisateur), la mise en relation des actions, quant à elle, rassemble celles déjà reconnues pour reproduire le scénario d'une vidéo. La deuxième contribution de notre travail est l'implantation et la validation d'un système pour l'interprétation des événements. Ce dernier consiste en l'extraction et l'interprétation des actions pour former un dictionnaire qui décrit efficacement et automatiquement, sans appel à l'intervention humaine, le contenu d'une vidéo.

Ce mémoire est organisé en 3 chapitres. Le premier, un état de l'art sur l'analyse des

actions humaines en général, et sur leur reconnaissance en particulier. Dans le chapitre 2, le modèle de reconnaissance d'actions humaines proposé et son expérimentation sont présentés. Le chapitre 3 aborde en détail la construction du dictionnaire. Finalement, une conclusion est dressée pour présenter les résultats obtenus et aussi pour discuter des perspectives.

CHAPITRE 1

État de l'art

L'analyse d'actions humaines est un des sujets les plus traités au sein de la communauté de la vision par ordinateur. En partant du domaine de la vidéosurveillance, à la médecine et en passant par celui des jeux vidéo, des applications récentes, dites intelligentes, se basent sur une telle analyse. Cette dernière consiste à détecter, à suivre au cours du temps et à reconnaître les activités d'une ou plusieurs personnes dans une vidéo. Dans les nombreux travaux consacrés à ce sujet, les chercheurs adoptent plusieurs approches, se basant chacune sur différentes caractéristiques.

Dans ce chapitre, nous dressons un état de l'art de l'analyse d'actions humaines, et plus précisément celui de la reconnaissance. Pour cela, nous répondons aux questions suivantes :

- Qu'est-ce qu'une action humaine ?
- Quel est l'intérêt d'analyser une action humaine ?
- Quelles sont les étapes de l'analyse d'actions humaines ?
- Quels sont les travaux consacrés seulement à la reconnaissance d'actions humaines ?
- Comment faire de la reconnaissance d'actions humaines ?

1.1 Définition d'une action

Une action est un processus effectué par un ou plusieurs objets qui changent dans le temps [91]. Cet objet peut être entre autres une voiture, un être humain, un arbre ou un robot. La définition étant trop large, Polana et Nelson [69] séparent les actions en deux catégories :

La première contient les actions dites stationnaires. Cette catégorie d'actions ne provoque pas un changement macroscopique (apparence) de l'objet. Elle représente un objet temporellement mobile, sans aucun effet sur l'apparence visuelle. Par exemple, une vidéo représentant la rotation d'un anneau circulaire et de couleur uniforme, ou une vidéo d'une chute d'eau.

La deuxième catégorie est celle des actions non stationnaires. La majorité des actions appartient à cette dernière catégorie, comme la vidéo d'une foule de personnes qui passe dans la rue. Cela se justifie par le fait que l'apparence visuelle des objets change au cours du temps. De ce fait, deux sous-catégories se distinguent. Les actions périodiques et non périodiques. Une action est dite non périodique, quand celle-ci ne se répète pas à travers le temps. Les séquences vidéos d'une personne qui effectue une chute ou une voiture qui rentre en collision avec une autre sont toutes considérées comme des actions non stationnaires, mais surtout non périodiques. Cependant, si une action dure dans le temps, alors elle est appelée action périodique. Les exemples suivants reflètent parfaitement ce type d'action : une personne qui marche, un cheval qui galope ou un joueur de hockey qui patine. Le schéma (figure 1.1) illustre ces différentes catégories d'actions.

Vu l'intérêt particulier des chercheurs pour l'être humain, beaucoup de travaux se sont portés à la fois sur les actions humaines non stationnaires [66, 44, 91] et sur les périodiques [77, 29, 60]. Le fait de pouvoir reconnaître, analyser et interpréter les activités d'une personne est utile pour différents domaines, tels que la vidéosurveillance ou la recherche

de vidéos par contenu.

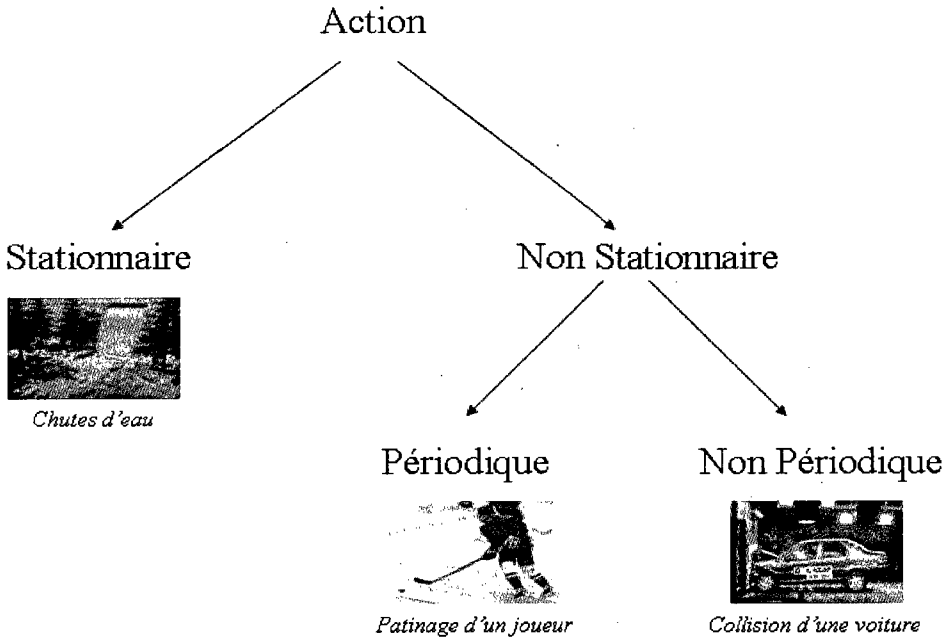


Figure 1.1 – Les catégories d'actions dans une vidéo.

1.2 Les domaines d'applications

Devant l'intérêt croissant pour l'analyse d'actions humaines, beaucoup d'applications sont parues. Certains travaux se sont limités au développement d'applications pour une action particulière. Par exemple, Petkovic *et al.* cherchent à reconnaître les revers dans un match de tennis [68]. D'autres, [60, 29] ont ciblé plusieurs actions humaines.

Nous regroupons les applications de l'analyse et de l'interprétation d'actions humaines en trois domaines. Le premier est l'analyse du mouvement. Ce domaine comprend les applications dont le but est de se focaliser sur une ou plusieurs parties du corps humain. Par exemple l'indexation, la recherche et l'analyse de vidéos de sport basées sur le contenu [52, 68]. La reconnaissance de problèmes orthopédiques [46, 61], dans le cadre d'études cliniques, fait partie aussi de ce domaine. Le deuxième domaine est celui de la surveillance vidéo. Ce dernier regroupe toutes les applications dont l'objectif est le suivi et le contrôle, au cours du temps, des actions d'une ou de plusieurs personnes. L'intérêt est de pouvoir sécuriser des lieux [23], contrôler l'accès à un site sensible en reconnaissant les visages par exemple [86] ou prévenir des accidents comme pour la surveillance des personnes âgées dans leurs résidences [27]. Certaines applications de *Marketing* font partie du domaine de la surveillance vidéo, comme celle qui nécessite de détecter et compter automatiquement des individus dans une vidéo [41]. Le troisième domaine est celui des interfaces avancées. Les applications appartenant à ce domaine visent à faciliter la communication entre des utilisateurs ou entre un utilisateur et une machine, comme la traduction des gestes (langage des signes) pour les sourds [81]. Cette communication peut concerner aussi un environnement virtuel, comme le cas des jeux vidéo interactifs [57].

Bien que ces applications soient nombreuses, l'analyse d'actions humaines, dans le domaine de la vision par ordinateur, se fait suivant des étapes bien précises.

1.3 Schéma général d'un système d'analyse d'actions humaines

L'analyse complète d'une vidéo contenant une ou plusieurs actions humaines suit une même approche. En effet, plusieurs études [84, 9] définissent trois étapes majeures pour

la réalisation de cette analyse. Le traitement commence par l'initialisation, ensuite se suit par le suivi du mouvement et se termine par la reconnaissance. Le schéma (figure 1.2) illustre l'enchaînement de ces étapes avec un aperçu de leurs composantes.

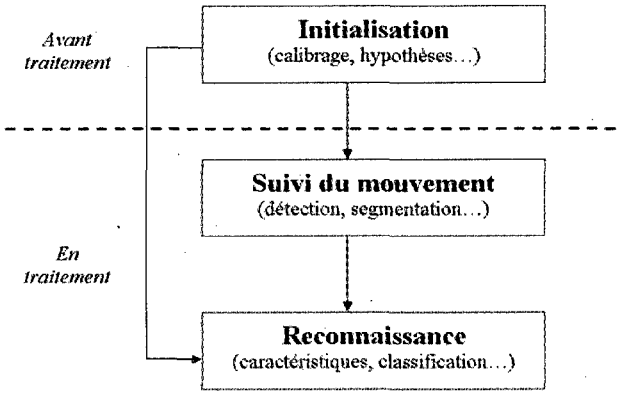


Figure 1.2 – Les étapes de l'analyse d'une action humaine

L'initialisation est une étape de prétraitement, qui a pour but la préparation des données nécessaires pour la suite du traitement. Elle peut consister à vérifier les hypothèses ou même effectuer le calibrage de la caméra. En général, cette étape nécessite une intervention manuelle de l'utilisateur.

L'étape du suivi du mouvement se compose, en général, de quatre sous-étapes. La première est la détection de mouvement. Cette dernière peut être vérifiée par un changement dans l'image de référence. La seconde, appelée segmentation, consiste à détecter l'objet mobile et le distinguer des autres, tout au long de la vidéo. Ici, l'objet représente un être humain en mouvement. La troisième sous-étape définit une représentation pour chaque

humain en mouvement. Cette représentation peut être sous forme d'un Blob [1] [20], d'une silhouette [35], d'une combinaison de caractéristiques spatio-temporelles [22], etc. Dans la quatrième sous-étape, une correspondance entre les représentations est effectuée, tout au long de la vidéo. Ceci permet le suivi, d'une manière continue, de l'être humain tout au long du mouvement. Dans certains travaux [35], une autre sous-étape s'ajoute. C'est l'identification d'une partie ou de la totalité du corps humain. La détection d'une partie du corps, comme les jambes [53], les bras [67] ou le corps humain en entier [54], est souvent utilisée pour faire de la reconnaissance d'actions humaines.

La reconnaissance est le résultat final de l'analyse d'une action humaine. Ce traitement est considéré comme une description sémantique de la vidéo (une personne qui marche, qui court, etc.). Deux catégories de reconnaissance peuvent être définies. La première, appelée statique, a pour but de reconnaître les postures d'une ou plusieurs personnes. Cette reconnaissance se fait en comparant l'information issue de chaque image (information spatiale) de l'action humaine avec celle issue de l'image préenregistrée. Certains travaux se contentent de comparer les silhouettes [36] pour déterminer la position d'une personne, si elle est par exemple debout, accroupie ou assise. Dans la deuxième catégorie, l'information temporelle, issue du mouvement, vient s'ajouter à l'information spatiale. Cette combinaison permet de différencier entre des actions semblables en apparence. Comme, une personne qui court et une autre qui marche rapidement. La reconnaissance se base sur un modèle de classification. Il suffit de comparer l'action (mouvement) détectée avec celles déjà définies dans l'étape d'apprentissage. Certains travaux se distinguent par la reconnaissance des actions sans information *a priori* [91].

Comme décrit précédemment, l'analyse d'actions humaines passe par les étapes d'initialisation, de suivi du mouvement et de reconnaissance. En réalité, la plupart des travaux

1. Blob signifie large objet binaire (Binary Large OBject). Un blob est une région de pixels connectés dans une image binaire ou en niveaux de gris. En général, le blob se base sur la similarité du mouvement (ou de la couleur aussi), ainsi que la proximité spatiale.

n'utilisent pas ces trois étapes ensemble. En effet, certains se sont focalisés sur le suivi du mouvement [10, 35, 36, 67], alors que d'autres sur la reconnaissance [59, 45, 91, 77, 29]. Donc, il est possible d'analyser des actions humaines en passant seulement par une étape de reconnaissance, sans considérer l'étape du suivi du mouvement. Dans ce cas, la reconnaissance nécessite un traitement préalable qui a pour but l'extraction d'informations (caractéristiques) pour représenter chaque action humaine.

1.4 Les données

Dans leurs travaux de reconnaissance d'actions humaines, les chercheurs utilisent des bases de données de vidéos. Chaque vidéo représente une personne [60] ou plusieurs [35] effectuant une action. Cette action peut représenter une activité quotidienne, par exemple le fait de marcher, courir, sauter ou s'assoir [45, 77, 59], ou une activité sportive, comme jouer au soccer ou pratiquer du tennis [66, 91, 20]. Pour le traitement, même si les vidéos sont en couleurs, elles sont souvent transformées en images à 256 niveaux de gris [77, 60, 91].

En fonction de l'objectif du travail, des hypothèses sur les vidéos sont émises. Dans le cadre de la reconnaissance d'actions humaines, ces hypothèses se portent sur le mouvement et sur l'apparence visuelle. Pour le mouvement, les plus fréquentes sont : le sujet reste dans le plan de la caméra, la caméra est fixe, pas d'occlusion, un mouvement lent et continu. Pour les hypothèses sur l'apparence visuelle, elles sont considérées selon l'environnement ou le sujet. Généralement, celles qui dépendent de l'environnement sont : une luminosité constante, un arrière-plan statique et uniforme. Alors que les hypothèses les plus utilisées selon le sujet sont : le point de départ connu, le sujet identifié. Le nombre d'hypothèses dépend de la complexité. Plus une vidéo est complexe, plus le nombre d'hypothèses est grand. Les hypothèses décrites sont détaillées dans le travail de Moeslund

et Granum [62]. Une fois les données établies, chaque action humaine est représentée par une ou plusieurs caractéristiques.

1.5 Représentation des vidéos

Pour faire de l'analyse d'actions humaines, et plus précisément pour la reconnaissance d'actions (courir, s'asseoir ou taper des mains), différentes caractéristiques sont extraites de la vidéo. Une caractéristique est un ensemble d'informations qui définissent une vidéo. Plusieurs caractéristiques peuvent être extraites d'une vidéo. Cependant, seulement certaines d'entre elles sont pertinentes. Une caractéristique est considérée pertinente, pour la reconnaissance d'actions humaines, quand elle distingue au mieux une action humaine des autres.

En 1973, Johansson [43] s'est intéressé à la caractérisation du mouvement humain. Pour son expérimentation, il a placé sur le corps d'un sujet, des cibles lumineuses au niveau de chaque articulation, ensuite il l'a filmé dans l'obscurité totale. Il a défini ainsi une caractéristique appelée PCL ou Points Caractéristiques Lumineux (*Moving Light Display*). Plus tard, Cutting et Kozlowski [26], en utilisant les PCL, se sont intéressés à distinguer les personnes selon leur démarche. Cette caractéristique montre qu'avec peu d'informations visuelles, il est possible de reconnaître quelques actions humaines.

Dans leur travail [15], Bobick et Davis utilisent l'Image de l'Historique du Mouvement (*Motion History Image* ou MHI) comme caractéristique. Le MHI regroupe le mouvement effectué au cours du temps en une seule image, où l'intensité d'un pixel dépend de la récence du mouvement. Cette récence est exprimée par l'Image de l'énergie du Mouvement (*Motion Energy Image* ou MEI). Par exemple, plus le mouvement d'un pixel est récent, plus l'intensité de ce pixel est grande et vice-versa. Cette caractéristique est illustrée par

la figure 1.3. L'inconvénient d'une telle caractéristique réside dans le choix du paramètre de durée. Meng *et al.* [60] utilisent le principe du MHI en introduisant la hiérarchie, appelée l'Histogramme Hiérarchique de l'Historique du Mouvement (*Hierarchical Motion History Histogramm* ou HMHH). Calculé de la même manière que le MHI, le HMHH est considéré selon plusieurs échelles de l'image. Cette dernière caractéristique est non seulement riche en information, mais aussi peu coûteuse en temps de calcul. En s'inspirant du MHI, Masoud et Papanikolopoulos [59] extrait des caractéristiques du mouvement humain directement de la vidéo. Ce dernier applique un filtre à Réponse Impulsionnelle Infinie (*Infinite Impulse Response*), pour chaque image. L'objectif est de pouvoir représenter le mouvement par sa récence, cette technique est appelée filtrage récursif.

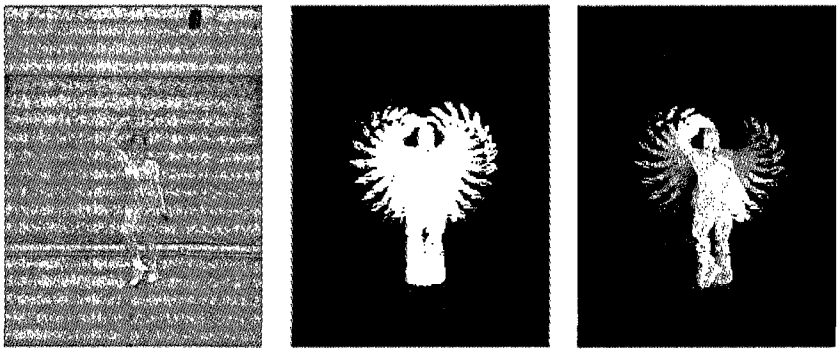


Figure 1.3 – Les caractéristiques MEI et MHI. (a) Image originale, (b) MEI et (c) MHI.

Pour représenter une action humaine dans une vidéo, le choix de points caractéristiques est une solution efficace. Dans ce sens, beaucoup de chercheurs optent pour des points contenant le maximum d'information sur le mouvement dans une vidéo, appelés points d'intérêts. Ces derniers, utilisés pour l'image [28], se sont généralisés à la vidéo [59]. Les

premiers à utiliser ces derniers points sont Schüldt *et al.* [77], dans le but de faire de la reconnaissance d'actions humaines. Avec la même approche que Harris et Stephens [37] pour l'extraction des points d'intérêts spatiaux, Laptev et Lindeberg [49] détectent les différents points, où l'information spatiale et temporelle locale connaît des variations significatives. Dollar *et al.* [29] proposent un détecteur de comportement et d'actions humaines (expression faciale et activités humaines) basé sur des points d'intérêts spatio-temporels (patches). Ce détecteur est un filtre temporel avec une paire de quadratures de filtres de Gabor à une seule dimension. Toujours pour la représentation d'une action par des points spatio-temporels, Kienzle *et al.* [45] s'inspirent du détecteur de Dollar *et al.* pour réaliser un détecteur basé sur un modèle du mouvement de l'œil humain. La figure 1.4 représente des points d'intérêts spatio-temporels, extraits d'une vidéo d'action humaine, selon différents détecteurs. Nous remarquons que la méthode de Kienzle *et al.*[45] produit le plus de points d'intérêts par rapport à celles de Dollar *et al.* [29] et de Schüldt *et al.* [77]. Cependant, la méthode [45] produits des points qui ne se trouvent pas sur l'objet en mouvement. L'étude de Niebles *et al.* [66] se distingue des autres par sa reconnaissance de plusieurs actions dans une même vidéo. Pour atteindre un tel objectif, chaque vidéo est représentée par une collection de rectangles, appelés Mots Spatio-Temporels (*Spatial-Temporal Words*), extraits à partir des points d'intérêts spatio-temporels. Chacun de ces rectangles appartient à une action déjà définie.

Pour Zelnik-Manor et Irani, une action est un objet temporel de longue durée. Dans leur travail [91], des caractéristiques spatio-temporelles locales sont extraites sous différentes échelles temporelles. L'objectif est de construire une pyramide temporelle (*temporal pyramid*) pour toute la vidéo. Ils estiment le gradient spatio-temporel local pour tous les points, à chaque échelle. Ils construisent après un histogramme de ce gradient normalisé. Finalement, une distribution multidimensionnelle (le nombre d'échelles multiplié par les dimensions spatio-temporelles) est estimée pour représenter chaque action.

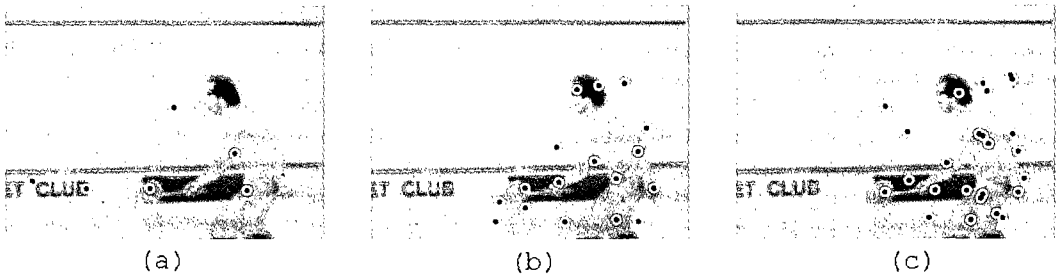


Figure 1.4 – Les points d'intérêts pour trois détecteurs spatio-temporels. (a) Schüldt *et al.* [77], (b) Dollar *et al.* [29], (c) Kienzle *et al.* [45].

D'autres caractéristiques, pertinentes aussi, sont utilisées pour reconnaître des actions humaines. Par exemple, Ke *et al.* [44] représentent les actions par des caractéristiques volumétriques spatio-temporelles. Ces dernières sont calculées à partir du flot optique horizontal et vertical de chaque point, tout au long de la vidéo.

1.6 Reconnaissance

Une fois les caractéristiques extraites de la vidéo, il reste à reconnaître l'action. Cette reconnaissance se fait par la comparaison de ses caractéristiques avec celles prédéfinies lors de l'initialisation. Pour cette fin, des modèles de classification sont développés. Le modèle de classification se compose en général d'une initialisation, d'un apprentissage et d'une comparaison. Lors de l'étape d'initialisation, des classes d'actions humaines sont définies *a priori*. L'apprentissage consiste à identifier ces classes, alors que la comparaison a pour résultat l'affectation d'une action humaine à sa classe d'appartenance.

Certains travaux optent pour une étape de réduction de données comme préalable à la classification. En effet, les chercheurs obtiennent ainsi une représentation moins re-

dondante, plus compacte et surtout plus discriminative, ce qui facilite la classification et réduit le temps de calcul. La méthode la plus utilisée est l'Analyse en Composantes Principales [44, 59, 48, 22].

Différents modèles de classification sont utilisés dans la littérature. Certains de ces modèles sont temporels, tels les modèles des Chaînes de Markov Cachées (*Hidden Markov Models* ou HMM). Pour la classification, HMM est un des modèles plus utilisés [48, 20]. L'objectif principal des HMM est de réaliser une association de données variant à travers le temps. En fait, les HMM sont des machines à états non déterministes qui, selon une entrée, passent d'un état à un autre par des probabilités de transition variées. Un des intérêts du HMM réside dans sa capacité à faire de la mise à jour des données d'apprentissage, tout au long du processus. Comme autre avantage, ce modèle permet d'ajouter facilement de nouvelles actions. Son long temps de calcul reste la faiblesse principale de ce modèle.

Les autres modèles, considérés comme non temporels, sont utilisés dans la reconnaissance d'actions humaines sous différentes formes. Comme le Séparateur à Vaste Marge (*Support Vector Machine* ou SVM) [77, 45, 60]. Malgré sa classification robuste, la fonction de décision est couteuse en temps de calcul. Cristianini et Shawe-Taylor [25] décrit parfaitement ce modèle. Une autre approche intéressante est celle des Réseaux de Neurones [45, 32]. Ce type de modèle a l'avantage d'être fonctionnel pour de grandes bases de données. D'autres travaux ont opté pour la méthode des K plus proches voisins (*k-nearest neighbors* ou Kppv) [29, 77]. Cette méthode nécessite de définir d'avance un nombre k de voisins. Pour chaque donnée en entrée, la classe, contenant le plus grand nombre de voisins trouvé parmi ces k voisins, est la classe d'appartenance. En général, le choix du nombre k influence beaucoup les résultats de la classification. Plus k est grand, plus il devient moins sensible au bruit, et moins il s'adapte aux petites bases d'apprentissage. Le modèle k-moyennes (*k-means*) est utilisé aussi comme un modèle non temporel [45]. Ce

modèle partitionne des données d'apprentissage en k classes distinctes, avec un nombre k fixé lors de l'initialisation. Après plusieurs itérations, un centre de gravité est trouvé pour chaque classe. La donnée est affectée alors à la classe du plus proche centre de gravité. Niebles *et al.* utilisent l'Analyse Sémantique Probabiliste (*Probabilistic Latent Semantic analysis* ou pLSA) comme une approche statistique pour construire un vocabulaire [66]. Cette méthode se base sur un apprentissage automatique des distributions de probabilités des mots spatio-temporels.

D'autres travaux se basent sur le calcul d'une distance pour classer leurs données. Le choix de la métrique joue un rôle important dans la classification d'actions humaines. Masoud et Papanikolopoulos [59] s'inspire de la distance de Hausdorff pour déterminer trois modèles de classification : Distance Minimum (DM), la Distance Moyenne Minimum (DMM) et la Moyenne des Distances Minimum (MDM). Dans [91], Zelnik-Manor et Irani utilise la distance χ^2 alors que Bobick et Davis [15] choisissent la distance de Mahalanobis pour comparer leurs caractéristiques.

1.7 Conclusion

Dans ce chapitre, un état de l'art sur l'analyse d'actions humaines en général, et la reconnaissance de ces actions en particulier, est dressé. Plusieurs travaux se sont focalisés sur le suivi du mouvement [35, 36, 67], alors que d'autres sur la reconnaissance [59, 45, 91, 77, 29]. Ces deux traitements, généralement successifs, peuvent être indépendants. Dans ce cas, la reconnaissance nécessite un traitement préalable qui a pour but l'extraction d'informations (caractéristiques) pour représenter chaque action humaine. Dans la littérature, plusieurs caractéristiques ont été utilisées. Elles peuvent être sous forme d'une image du mouvement ou de points d'intérêts. Sauf qu'il est difficile de déterminer la caractéristique pertinente, celle qui distingue le mieux une action humaine.

Après l'extraction de la caractéristique, une vidéo est affectée, selon un modèle de classification, à la catégorie d'action correspondante. Dans la suite de ce mémoire, nous allons proposer notre approche pour la reconnaissance d'actions humaines. En basant sur une nouvelle caractéristique, nous utilisons différents modèles de classification. L'objectif est de mettre en valeur l'apport de notre caractéristique, ainsi que les performances de chacun des modèles utilisés.

CHAPITRE 2

Approche pour la reconnaissance d'actions humaines

L'objectif de notre travail est la reconnaissance d'actions humaines. Nous procédons par étapes dans le but d'avoir des résultats comparables par rapport à l'existant. D'abord, nous cherchons les caractéristiques qui représentent au mieux nos actions vidéo. Puis, nous procédons à une réduction pertinente des données, ce qui nous permet de réduire le volume des données à traiter, et enfin nous procédons à la classification selon le modèle le plus adapté. La figure 2.1 représente un schéma récapitulatif de notre approche pour la reconnaissance d'actions humaines.

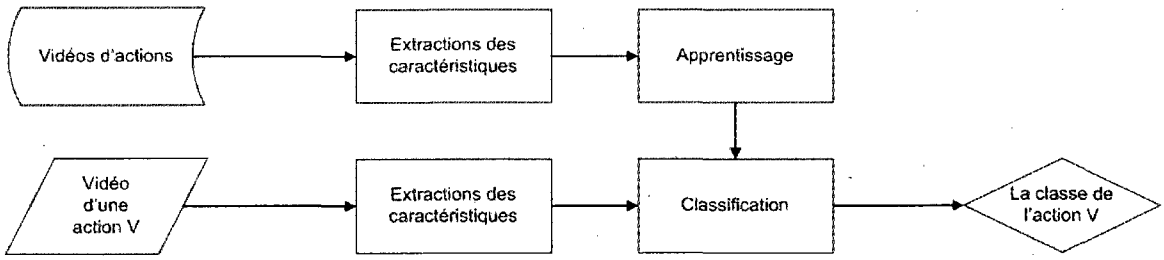


Figure 2.1 – Schéma de l'approche pour la reconnaissance d'actions humaines.

2.1 Extraction des caractéristiques

Le choix des caractéristiques dans notre modèle découle des méthodes de reconnaissance d'actions humaines (Section 1.5). Ces méthodes reposent généralement sur une caractéristique spatio-temporelle. Que cela soit les points d'intérêts spatio-temporels de Laptev et Lindeberg [49] ou les patches de Dollar *et al.* [29], ce genre de caractéristiques s'est avéré plus efficace pour la reconnaissance d'actions humaines que les modèles basés sur le flux optique ou les caractéristiques de suivi. Parmi les limites de ces dernières caractéristiques, il y a les problèmes d'occlusion ou de formes d'objets non rigides [49, 29, 91, 66].

Une caractéristique spatio-temporelle intègre une information spatiale indiquant la forme de l'objet et une information temporelle qui indique un changement temporel. Les deux informations combinées apportent alors une information sur un objet en mouvement. Parmi les caractéristiques spatio-temporelles, les points d'intérêts spatio-temporels de Laptev et Lindeberg [49] et les patches de Dollar *et al.* [29] sont les plus connues, d'ailleurs d'autres caractéristiques en découlent. La limite de la première réside dans le peu d'informations générées [29], alors que le temps de calcul de la deuxième représente sa limitation principale [66]. Dans notre travail, nous cherchons à remédier à ces limites

par le choix de deux caractéristiques au lieu d'une seule. Notre choix est motivé par le taux élevé d'information générée que peuvent avoir deux caractéristiques combinées, tout en ayant un temps d'exécution raisonnable.

2.1.1 Les points d'intérêts spatio-temporels

Dans plusieurs travaux de reconnaissance d'actions humaines (section 1.5), nous remarquons que les points d'intérêts sont les plus utilisés et cela pour les avantages qu'ils offrent : robustesse aux occlusions et aux problèmes d'ouverture, présence sur presque toutes les images, etc. Un point d'intérêt spatial se définit comme une discontinuité, dans deux dimensions, de la fonction d'intensité ou de ses dérivés. Laptev et Lindeberg ont généralisé cette définition pour atteindre trois dimensions [49]. Les points d'intérêts sont utilisés dans plusieurs domaines, dont la reconnaissance, la segmentation et la mise en correspondance. Différents détecteurs de points d'intérêts existent, notamment le détecteur de Harris, détecteur de Moravec, SIFT. Dans notre travail, nous nous intéressons au détecteur SIFT (*Scale Invariant Features Transforms*) [28] pour ses propriétés intéressantes. Nous décrivons en ce qui suit ce détecteur.

Soit I une image donnée et $L(x, y, \sigma)$ l'espace échelle défini par :

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \tag{2.1}$$

avec

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}, \tag{2.2}$$

où σ est l'échelle.

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) && (2.3) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) && (2.4) \end{aligned}$$

La différence de gaussienne LoG est approximée en utilisant l'équation de la diffusion de la chaleur tel que :

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2\nabla^2 G, \tag{2.5}$$

Par conséquent,

$$D(x, y, \sigma) \equiv (k-1)\sigma^2\nabla^2 G * I(x, y), \tag{2.6}$$

Il s'agit maintenant de localiser les extrémums de $D(x, y, k\sigma)$. Pour cela, chaque point est comparé avec ses huit premiers voisins, ainsi que les neuf autres voisins de l'échelle d'après et l'échelle d'avant. Ce point est considéré maximum local, si sa valeur est plus grande que celles de ses voisins. De la même manière, les minimums locaux sont obtenus. Cette méthode pose le problème des faux extrémums, en supprimant ceux qui possèdent une faible valeur de $D(x, y, k\sigma)$. Comme le soulève D.G. Lowe [28], deux extrémums peuvent exister pour chaque côté d'une ellipse. Il est donc nécessaire d'affiner cette recherche d'extrémums. Pour ce faire, l'approche utilisée est celle du développement de Taylor de la fonction multi-échelle $D(x, y, \sigma)$. Pour un point, cette fonction est représentée comme suit :

$$D(x) = D + \frac{\delta D^T}{\delta x}x + \frac{1}{2}x^T\frac{\delta^2 D^T}{\delta x^2}x \tag{2.7}$$

avec D la dérivée à un point $x = (x, y, \sigma)^T$. Lorsque, la dérivée de cette fonction est égale à zéro, l'extrémum $\hat{x}$ est déterminé, tel que $\hat{x} = -\frac{\delta^2 D^{-1}}{\delta^2 x^2}\frac{\delta D}{\delta x}$. La fonction $D(\hat{x})$ est utilisée pour rejeter les extrémums avec un contraste faible. Il suffit pour cela de substituer $\hat{x}$ dans l'équation 2.7 et vérifier si $D(\hat{x})$ est supérieur à un certain seuil r.

$$D(\hat{x}) = D + \frac{1}{2}\frac{\delta D^T}{\delta x}\hat{x} \tag{2.8}$$

Cela reste insuffisant, puisque des points appartenant au contour, ayant un contraste fort, ne sont pas réellement des points d'intérêts. Pour remédier à ce problème, une solution, en exploitant la courbure, a été proposée par D.G. Lowe [28]. Cette solution consiste

à éliminer les points d'intérêts ayant une large courbure principale dans la direction du contour et une petite dans la direction perpendiculaire. La matrice H (matrice Hessienne) est utilisée pour le calcul des courbures. Les valeurs propres de H sont proportionnelles à la courbure principale de D. En considérant ces remarques, il suffit de chercher les valeurs propres de la matrice Hessienne 2×2 aux coordonnées et à l'échelle du point d'intérêts, afin de savoir si l'extrémum est mal défini ou pas.

$$H = \begin{pmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{pmatrix}$$

L'approche de Harris et Stephens [37] montre qu'il n'est pas nécessaire de chercher les valeurs propres, mais seulement leur ratio. Soit α la valeur propre ayant la plus grande amplitude et β celle ayant la plus petite. Il est possible de calculer la somme des valeurs propres par la trace (Tr) de H, et le produit par le calcul du déterminant (Det) de H tel que :

$$TR(H) = D_{xx} + D_{yy} = \alpha + \beta \tag{2.9}$$

$$Det(H) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta, \tag{2.10}$$

Soit r, le rapport entre α et β tel que $\alpha = r\beta$ alors

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r}, \tag{2.11}$$

La quantité $\frac{(r+1)^2}{r}$ diminue lorsque les deux valeurs propres sont égales, et elle augmente quand r augmente. Pour vérifier si le ratio de la courbure principale est plus petit qu'un certain seuil r, il suffit de vérifier :

$$\frac{TR(H)^2}{Det(H)} < \frac{(r+1)^2}{r} \tag{2.12}$$

Notons que $r = 10$ définit expérimentalement par [28].

Pour chaque point d'intérêt, un vecteur de caractéristiques est calculé tel que ce vecteur préserve l'invariance à la rotation. D.G. Lowe [28] a procédé au calcul de l'orientation $\theta(x, y)$ et l'amplitude $m(x, y)$ du gradient, pour chaque point de L, tel que :

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad (2.13)$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))) \quad (2.14)$$

Un histogramme d'orientations est calculé, alors, dans une région autour du point d'intérêt trouvé. Chaque région est formée de $16 * 16$ éléments divisés en quatre sous régions. Chaque sous région contient quatre histogrammes d'orientations. Chaque histogramme est échantillonné en huit *bins* selon les orientations. Les amplitudes sont accumulées dans chaque *bins*. D'où un vecteur de caractéristiques de $4 * 4 * 8 = 128$ données.

Ce détecteur donne des points étant représentés chacun par un vecteur de caractéristiques invariant au changement de rotation, de translation, d'illumination et au petit changement de point de vue. Pour notre travail, nous adaptons ce détecteur en fonction de nos besoins. Malgré ses avantages, ce détecteur présente certaines limitations. Premièrement, c'est un détecteur spatial. Deuxièmement, appliqué à une image, ce dernier fournit beaucoup d'informations (un grand nombre de points, avec pour chaque point un vecteur de 128 données). Pour une vidéo, cette quantité d'information devient difficile à gérer. Troisièmement, ce détecteur génère des points qui ne sont pas très utiles pour notre travail. Comme le cas des points qui se trouvent dans le fond des images ou dans un objet fixe de la vidéo. La lenteur d'exécution du détecteur reste sa dernière limitation, elle est due au grand nombre de convolutions à effectuer pour la détection de points. L'exécutable fourni par D.G. Lowe [28] prend à peu près six secondes pour une image de 200×300 pixels. Pour le traitement d'une vidéo de 1 minute échantillonnée à 20 images par seconde, ce détecteur prend 2 heures ($60 * 20 * 6 = 7200$ secondes).

Maintenant, nous cherchons ce que nous appelons les zones d'intérêts. Chaque zone est une région de l'image où un mouvement est effectué. Pour détecter cette région, nous calculons le contour spatio-temporel (défini dans la section suivante). Nous agrandissons la région trouvée, pour être sûr que l'objet en mouvement soit considéré. Dans nos expériences, nous augmentons la taille de cette région de quinze pixels dans les deux sens de l'image. Donc, cette région de l'image contient de l'information sur le mouvement dans la vidéo. Après avoir détecté les zones d'intérêts dans la vidéo, il suffit de chercher les points d'intérêts, selon le détecteur SIFT, se trouvant dans chaque zone. Ces points d'intérêts sont alors des points d'intérêts spatiaux dans une zone de changement temporel (zone d'intérêt). La figure 2.2 page 27 illustre ces points d'intérêts contenus dans une zone d'intérêt.

Dans la suite, ces points sont appelés Points d'Intérêts Spatio-Temporel (PIST). Notre détecteur de points d'intérêts est nommé alors SIFT spatio-temporel. Ce détecteur a l'avantage de donner des points d'intérêts d'une quantité suffisante pour caractériser seulement le mouvement dans une vidéo. Nous obtenons un temps d'exécution plus rapide que le SIFT, tout en gardant les mêmes avantages (invariance au changement de rotation, de translation, d'illumination et au petit changement de point de vu).

2.1.2 Le contour spatio-temporel

Après avoir extrait les points d'intérêts spatio-temporels de la vidéo, nous remarquons que certains points n'appartiennent pas au mouvement de l'objet. Comme illustré par la figure 2.2, nous remarquons des points sur la poitrine du sujet, alors que celle-ci ne bouge pas. Pour avoir plus de précision sur le mouvement, nous choisissons, comme deuxième caractéristique, le contour spatio-temporel. Le contour englobe tout le mouvement d'un objet. En plus, nous l'avons déjà calculé précédemment, lors de l'extraction des zones

d'intérêts, ce qui nous garantit un gain considérable en temps de calcul.

Un contour spatial dans une image $I(x, y)$ est une ligne sur laquelle se produit les plus grandes variations de I. Soit G le gradient de I défini par :

$$G = \vec{\nabla} I = \begin{pmatrix} \frac{\partial I}{\partial x} \\ \\ \frac{\partial I}{\partial y} \end{pmatrix} \tag{2.15}$$

Le module et le vecteur directeur du gradient de I sont donnés par :

$$|G| = \left| \vec{\nabla}\, I \right| = \left[\left(\frac{\partial I}{\partial x} \right)^2 + \left(\frac{\partial I}{\partial y} \right)^2 \right] \tag{2.16}$$

$$\vec{g} = \frac{\vec{\nabla} I}{\left| \vec{\nabla} I \right|} \tag{2.17}$$

Un point appartenant au contour spatial est trouvé lorsque le module de son gradient est supérieur à un certain seuil. Pour inclure l'axe du temps, nous généralisons ce contour en nous inspirant des travaux de Laptev et Lindeberg [49]. Ce nouveau contour est appelé Contour Spatio-Temporel (CST). D'après [49], nous pouvons construire pour une séquence d'images $V : \Re^2 \times \Re_+ \rightarrow \Re$, un espace d'échelle linéaire $L : \Re^2 \times \Re \times \Re^2_+ \longmapsto \Re$ en appliquant une convolution à V par un noyau gaussien spatio-temporel de variance spatiale σ_l^2 et de variance temporelle τ_l^2.

Étant donnée une séquence V de N images telle que :

$$V(x, y, t) = I_1(x, y), I_2(x, y), ..., I_N(x, y) \tag{2.18}$$

et un noyau gaussien spatio-temporel défini par :

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}}\, exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2) \tag{2.19}$$

alors,

$$L(\cdot; \sigma_l^2, \tau_l^2) = g(\cdot; \sigma_l^2, \tau_l^2) * V(\cdot) \tag{2.20}$$

À partir de ce modèle, nous cherchons le gradient spatio-temporel F en appliquant une convolution à V par les dérivées partielles de la fonction du noyau gaussien g tel que :

$$\vec{F}=\vec{\nabla}V=\begin{bmatrix}\frac{\partial V}{\partial x}\\ \\ \frac{\partial V}{\partial y}\\ \\ \frac{\partial V}{\partial t}\end{bmatrix} \tag{2.21}$$

avec

$$\frac{\partial V}{\partial x} = V_x(x,y,t) = g_x * V(x,y,t) \tag{2.22}$$

$$\frac{\partial V}{\partial y} = V_y(x,y,t) = g_y * V(x,y,t) \tag{2.23}$$

$$\frac{\partial V}{\partial t} = V_t(x,y,t) = g_t * V(x,y,t) \tag{2.24}$$

Pour détecter le contour, nous calculons le module du gradient $|F|$ donnée par :

$$|F| = \left|\vec{\nabla}\, V\right| = \left[\left(\frac{\partial V}{\partial x}\right)^2 + \left(\frac{\partial V}{\partial y}\right)^2 + \left(\frac{\partial V}{\partial t}\right)^2\right] \tag{2.25}$$

et nous vérifions si ce module est supérieur à un certain seuil r. Nous calculons aussi l'orientation du gradient. Comme le gradient spatio-temporel est en 3D, son orientation doit comprendre alors deux angles, θ qui est l'angle spatial et ρ l'angle temporel par rapport à x tel que :

$$\theta = \arctan(F_y/F_x) \; et \; \rho = \arctan(F_x/F_t) \tag{2.26}$$

Ces orientations sont donc appelées Contour Spatio-temporel (CST). Cette caractéristique CST est illustrée par la figure 2.3 page 28.

2.1.3 Algorithme et résultats expérimentaux

En résumé, l'algorithme d'extraction des caractéristiques prend en entrée une vidéo V, et génère en sortie deux caractéristiques CST (Contour Spatio-Temporel) et PIST (Points d'Intérêts Spatio-temporel). D'abord, pour chaque image i de V, les trois dérivées partielles en i par rapport à x, y et t sont calculées selon l'équation 2.21. Le module du gradient est calculé selon l'équation 2.25. Nous vérifions pour chaque point, s'il est supérieur à un certain seuil r. Ainsi, le Contour Spatio-Temporel (CST) est obtenu. Les deux vecteurs directeurs du gradient spatio-temporel sont ensuite calculés à chaque point du contour, selon l'équation 2.26. La zone qui englobe le CST définit alors notre zone d'intérêt, dont la dimension dépend du CST obtenu. Après, les différences d'images sont calculées à différentes échelles pour chaque zone, selon l'équation 2.4. Nous trouvons ensuite les extrémums locaux de $D(x, y, k\sigma)$. À partir de ces derniers, nous éliminons tous les points ayant un contraste faible, en substituant les points ayant $D(\hat{x})$ (équation 2.8) plus petit qu'un seuil donné. Les points avec un contraste élevé ayant une large courbure sont aussi enlevés, en utilisant le ratio des valeurs propres de la matrice Hessiennne (équation 2.12). Les points restants sont alors les PIST. Le vecteur directeur et l'amplitude de chaque PIST sont calculés selon l'équation 2.13 et 2.14. Finalement, le vecteur de caractéristiques de 128 données est construit pour chaque PIST.

Nous avons expérimenté les caractéristiques CST et PIST sur 600 vidéos d'actions humaines. Ces vidéos sont échantillonnées à 25 images à niveaux de gris par secondes, avec une résolution de 160×120. Pour ces tests, nous avons utilisé un ordinateur muni d'un processeur T5750 Intel Core 2 Duo, cadensé à 2*2.00 GHz et muni de 3 Go de mémoire vive. Pour la caractéristique CST, nous obtenons un temps de traitement moyen de 28 secondes pour une vidéo contenant 500 images. Ceci donne un taux de traitement moyen de 18 images/seconde. Nous constatons alors que l'extraction du CST est rapide. Pour extraire les PIST, notre algorithme a besoin, en moyenne, de 10 minutes pour traiter

une vidéo de 500 images. Donc le taux de traitement moyen est de $0,8$ image/seconde. Nous remarquons que ce temps obtenu est court aussi. En comparaison avec l'exécutable de D.G. Lowe [28], pour une même taille d'image, nous obtenons un taux moyen de $0,5$ image/seconde. Ceci confirme la rapidité de notre algorithme. Comme mentionné précédemment, les PIST nécessitent l'extraction du CST afin d'obtenir les zones d'intérêts. Donc le temps de traitement obtenu pour les PIST inclut celui de l'extraction du CST. Nous appelons ces deux caractéristiques combinées l'information CSST (Contour et Sift Spatio-Temporel).

La figure 2.2 présente trois images décrivant une personne qui bouge les bras. Comme illustré par cette figure, certains points, détectés parmi les PIST, se trouvent sur l'abdomen du sujet. Or, aucun mouvement n'est effectué dans cette partie de l'image. Donc les fausses détections sont considérées comme des limitations pour les PIST. Dans le cas de variations du zoom dans une vidéo, la caractéristique du CST confond le mouvement de la caméra avec le mouvement de l'action. Ce que nous considérons comme une des limitations du CST. La figure 2.3, illustre les résultats obtenus pour la caractéristique CST.

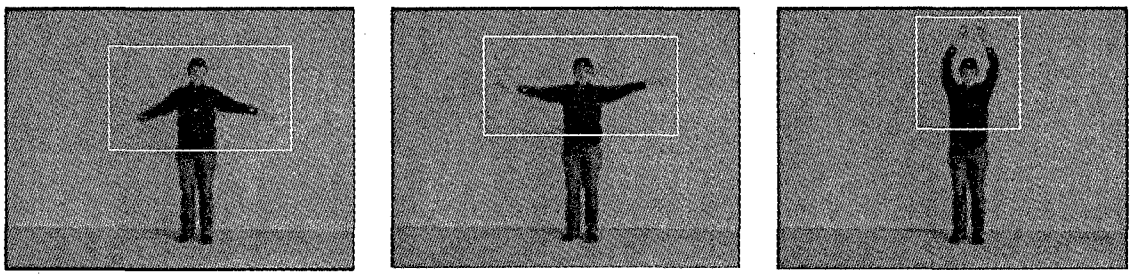


Figure 2.2 – Les points d'intérêts dans une zone d'intérêt

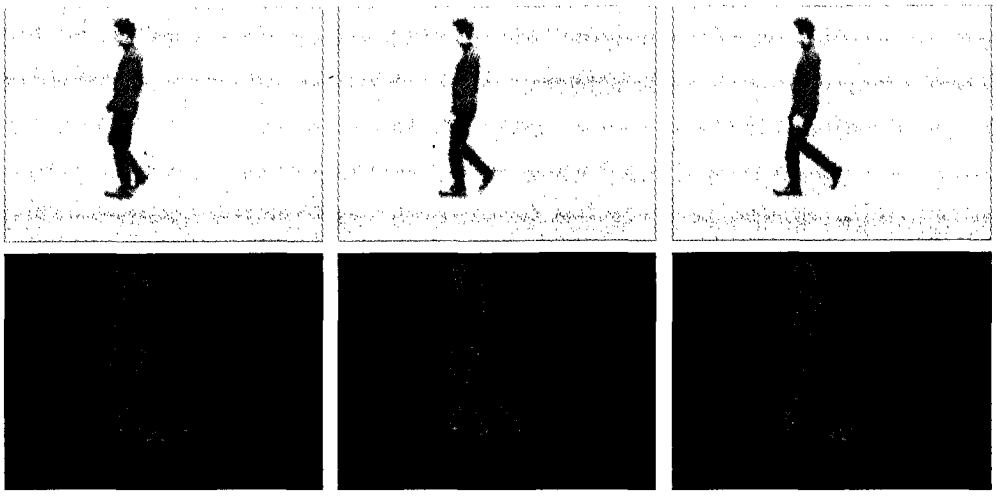


Figure 2.3 – Le contour spatio-temporel pour $\sigma_l^2 = 1.2$ et $\tau_l^2 = 1.0$.

2.2 Réduction des données

Notre objectif est d'obtenir des caractéristiques en quantité suffisante et significatives. Les caractéristiques spatio-temporelles, obtenues dans la section précédente, sont appelées la caractéristique CSST (Contour et Sift Spatio-Temporel). Pour chaque vidéo un grand nombre de données CSST sont générées. En effet, pour chaque image I de la vidéo V, le SIFT Spatio-Temporel fournit un nombre de points d'intérêts N_i. Le nombre total des PIST (Points d'intérêts Spatio-Temporel) pour une vidéo de m images est :

$$N_m = \sum_{i=0}^{m} N_i \tag{2.27}$$

Chaque PIST correspond à un vecteur de 128 données. Alors, pour une vidéo de $m = 200$ images et de $N_i = 50$ points d'intérêts en moyenne, nous obtenons 10000 vecteurs de 128 dimensions. Ceci représente une grande quantité de données à gérer. Dans la plupart des modèles de classification, que ça soit le SVM, le Kppv ou la Régression logistique, chaque donnée en entrée (vidéo) est représentée par un vecteur de données décrivant une

caractéristique. En général, ces modèles considèrent un nombre limité de caractéristiques. Pour cette raison, il est impératif de diminuer le grand nombre de données obtenues par le SIFT spatio-temporel. Nous utilisons, à cette fin, l'Analyse en Composantes Principales (ACP). Soit $\vec{p}_1$,...$\vec{p}_{N_m}$ des PIST d'une vidéo V, et M la matrice des vecteurs de ces PIST, centrée et réduite.

$$M = \begin{array}{cc} & \\ I_1 & \vec{p}_1 \\ & \vdots \\ & \vec{p}_{N_0} \\ I_2 & \vec{p}_1 \\ & \vdots \\ & \vec{p}_{N_1} \\ I_i & \vec{p}_1 \\ & \vdots \\ & \vec{p}_{N_i} \\ I_m & \vec{p}_1 \\ & \vdots \\ & \vec{p}_{N_m} \end{array} \begin{pmatrix} a_{1,1} & \dots & a_{1,128} \\ \vdots & & \vdots \\ a_{N_0,1} & \dots & a_{N_0,128} \\ a_{1,1} & \dots & a_{1,128} \\ \vdots & & \vdots \\ a_{N_1,1} & \dots & a_{N_1,128} \\ a_{1,1} & \dots & a_{1,128} \\ \vdots & & \vdots \\ a_{N_i,1} & \dots & a_{N_i,128} \\ a_{1,1} & \dots & a_{1,128} \\ \vdots & & \vdots \\ a_{N_m,1} & \dots & a_{N_m,128} \end{pmatrix}$$

La matrice M_{vc} de variance-covariance est donnée par :

$$M_{vc} = \frac{1}{N_v} M^t M \tag{2.28}$$

Le principe de l'ACP est de trouver un axe u, issu d'une combinaison linéaire de nos vecteurs de données, tel que la variance du nuage autour de cet axe soit maximal. Nous appliquons l'ACP sur nos données pour réduire la dimension de notre espace. Après expérimentation, nous remarquons qu'en moyenne, la première valeur propre associée au premier vecteur propre est de 28.88% avec un écart-type 3.25%, et atteint 58.1% pour les 5 premières avec un écart-type de 2.23%. Ces résultats sont calculés pour un échantillon de 182 vidéos parmis les vidéos de la base de données décrite dans la section

suivante. En se limitant aux cinq premiers vecteurs propres, le nombre de données diminue significativement. Mais, il en reste que c'est toujours un grand nombre pour les modèles de classification.

D'après l'ACP, le premier vecteur propre est la meilleure représentation des données en une dimension (1D). Partant du fait que notre premier vecteur propre décrit en moyen 28.88% de l'inertie totale des vidéos, ce qui est considérable pour 128 dimensions en total, et sachant que c'est la meilleure représentation en (1D), nous choisissons, alors, de représenter nos données selon le premier vecteur propre de dimension 128. Ceci réduit considérablement la représentation des données.

Pour le CST, les vecteurs directeurs sont utilisés comme caractéristique à part. Pour chaque image I, l'orientation temporelle génère un grand nombre de données, surtout si le mouvement est rapide. Par conséquent, une étape de réduction de données s'impose. Nous créons pour chaque image I de la vidéo V, un vecteur d'orientations de dimension l, en effectuant un échantillonement d'un bin égal à $\frac{\pi}{l}$. Nous avons expérimentalement les meilleurs résultats de classification pour $l = 32$. Ce vecteur se compose du nombre d'orientations par bin de longueur égale à $\frac{\pi}{32}$ pour chaque image, (où les orientations varient de 0 à π). Par exemple, pour une vidéo contenant 300 images, il y a 300 vecteurs de 32 orientations. À la fin de la réduction des données par l'ACP pour cette caractéristique, il en résulte deux vecteurs de 32 valeurs chaque, un vecteur pour les orientations spatiales et un pour les orientations temporelles. Ceci est illustré par le tableau 2.1

3	0	0	....	22	1	9	0
$[0, \frac{\pi}{32}[$	$[\frac{\pi}{32}, \frac{\pi}{16}[$	$[\frac{\pi}{16}, \frac{3*\pi}{32}[$	....	$[\frac{28*\pi}{32}, \frac{29*\pi}{32}[$	$[\frac{29*\pi}{32}, \frac{30*\pi}{32}[$	$[\frac{30*\pi}{32}, \frac{31*\pi}{32}[$	$[\frac{31*\pi}{32}, \pi[$

Tableau 2.1 – Exemple de vecteur d'orientations du gradient

En résumé, notre algorithme de réduction de données prend en entrée tous les PIST de la vidéo V et le vecteur directeur de chaque point appartenant au CST de V. Rappelons que les PIST sont calculés séparément pour chaque zone d'intérêt. Nous obtenons en sortie un seul vecteur pour tous les PIST et un histogramme des orientations pour chaque zone d'intérêt de V. Nous commençons par rassembler tous les PIST dans une matrice M. Cette dernière est ensuite réduite, centrée à l'aide de l'ACP. Notons que les points, qui se répètent, sont éliminés, donc ils ne sont pas considérés par l'ACP. La dimension d'une telle matrice M est de 128* le nombre des PIST obtenus. Selon l'équation 2.28, la matrice de variance-covariance M_{vc} est calculée. De celle-ci, nous obtenons les vecteurs propres et les valeurs propres de M. Au final, nous choisissons le premier vecteur propre pour qu'il soit le vecteur représentatif des PIST. Pour les orientations, l'algorithme consiste d'abord à construire un vecteur de 32 *bins*, pour chaque zone d'intérêt. Ensuite, le nombre d'orientations variant dans un *bin* de longueur $\pi/32$ est calculé, tel que les angles du vecteur varient entre 0 et π.

2.3 Apprentissage et classification des actions humaines

Après l'extraction des caractéristiques et leur réduction, nous cherchons une méthode pour la classification des actions humaines présentes dans les vidéos. Inspirés des travaux antérieurs, nous choisissons d'abord un modèle de classification simple. Ce modèle est la classification par les K plus proches voisins (Kppv). Et nous expérimentons, ensuite, un nouveau modèle de classification des actions humaines, appelé Modèle Bayésien pour la Régression Logistique.

2.3.1 Classification par les K plus proche voisins (Kppv)

Soit $y_1, \cdots, y_c$ des classes d'actions humaines comme marcher, courir, applaudir, avec c le nombre total d'actions. Comme illustrée dans la figure 2.4, une action est formée de plusieurs zones d'intérêts non nécessairement connexes. À chaque action est associée alors une information CSST. Cette information est représentée par un ensemble de vecteurs, qui nous permettent d'affecter une vidéo V à une classe y_i. Pour cela, il est indispensable d'avoir un modèle de classification. Un des plus utilisés dans la littérature (Section 1.6) est le Kppv ou les K plus proches voisins.

Avant d'effectuer la classification, ce modèle passe par une étape d'apprentissage. En effet, un ensemble d'apprentissages T se constitue des informations CSST des vidéos d'apprentissage. T est divisé en sous-ensembles, représentant chacun une classe d'actions humaines y_i. Donc chaque vidéo de l'ensemble d'apprentissages est étiquetée selon sa classe d'appartenance. Pour classer une vidéo V en entrée, le modèle calcule d'abord la distance $d(\cdot, \cdot)$ (euclidienne, cosinus...) entre l'information CSST de V et celles de toutes les vidéos de l'ensemble T. Ensuite, le Kppv effectue un tri croissant des distances obtenues. En vérifiant l'étiquette qui se répète pour les k plus proches voisins, le modèle attribue une classe d'appartenance à V selon cette étiquette.

Figure 2.4 – Les zones d'intérêts.

Sachant que le CSST se compose de deux informations, nous choisissons d'abord de traiter par le Kppv chacune de ces informations à part, et d'additionner ensuite les résultats obtenus selon une pondération. En effet, pour la caractéristique PIST, une vidéo V est représentée par un seul vecteur U, contrairement au CST qui en résulte un vecteur pour chaque zone d'intérêt Z_j. Donc pour une vidéo V, nous cherchons par le Kppv la vidéo ayant la plus petite distance $d(\cdot,\cdot)$ avec U. Pour le CST, nous calculons une distance qui sépare chaque vidéo V d'une vidéo d'apprentissage V_T de T. Cette distance est une moyenne des distances entre chaque zone d'intérêt Z_j de V et la zone correspondante Z_{T_i}. Cette dernière se définit comme une zone de la vidéo d'apprentissage V_T, ayant la plus petite distance avec Z_j.

Pour le Kppv, Qian *et al.* [70] montrent que la distance euclidienne et la distance cosinus sont similaires et donnent de bons résultats, lorsque le nombre de voisins k est grand. Toujours selon ces auteurs, dans le cas d'une combinaison de caractéristiques, la distance cosinus devient plus performante que l'euclidienne, avec un temps d'exécution plus court. Lorsque le nombre de voisins k est petit, la classification par ces deux distances est variante. Sachant qu'il existe peu de bases de données avec une grande quantité de vidéos d'actions humaines, un grand échantillon d'apprentissage devient difficile à former. Nous devons pour ça étudier ces deux distances. La distance euclidienne se définit pour un vecteur de n dimensions, pour V (la donnée à classer) et V_T (une donnée appartenant à l'échantillon d'apprentissage), par :

$$d(.,.) = \sqrt{\sum_{i=1}^{n}(V_i - V_{T_i})^2} \tag{2.29}$$

et la distance cosinus par :

$$d(.,.) = \cos(V, V_T) = \frac{V.V_T}{|V||V_T|} \tag{2.30}$$

2.3.2 Algorithme de classification par Kppv

Notre algorithme pour le Kppv a en entrée un ensemble d'apprentissages T_S des premiers vecteurs propres pour les PIST, un ensemble d'apprentissages T_H de vecteurs d'histogrammes d'orientations du CST, une vidéo V d'une action humaine et un nombre de voisins k. En sortie, nous obtenons la classe d'actions correspondante à la vidéo V. Nous commençons par l'extraction des caractéristiques, pour appliquer après l'algorithme de réduction de données pour obtenir le vecteur propre V_S et l'ensemble de vecteurs V_H. Nous formons ensuite l'ensemble $E_S(V_S, T_S, d(.,.), k)$ des k plus proches distances calculées pour V_S par rapport à T_S. La distance $d(.,.)$ est soit euclidienne ou cosinus. De la même manière, nous formons l'ensemble $E_H(V_H, T_H, d(.,.), k)$ des k plus proches distances calculées pour V_H par rapport à T_H. Pour construire un tel ensemble, nous calculons la distance entre V_H^i et t_H^j pour chaque vecteur d'histogramme V_H^i de V_H et t_H^j de t_H (une vidéo appartenant à T_H). Cette dernière étape est répétée chaque vecteur t_H^j. Nous retenons ainsi la distance d_{min}, étant la plus petite par rapport à V_H^i. Nous répétons aussi ces trois dernières étapes pour le reste des vecteurs V_H. Une distance moyenne $d_{min}(t_H)$ de toutes les d_{min} est ainsi calculée. Nous obtenons donc, pour chaque élément t_H de T_H, une distance $d_{min}(t_H)$ par rapport à V_H. Nous calculons cette dernière distance pour le reste des éléments t_H de T_H. Pour chaque classe y_i, nous calculons le nombre d'éléments D_S dans $E_S(V_S, T_S, d(.,.), k)$, et le nombre d'éléments D_H, contenu dans $E_H(V_H, T_H, d(.,.), k)$. Nous combinons alors les deux nombres de voisins pour chaque classe y_i, avec une pondération α et β trouvée expérimentalement, tel que le nombre total $D_T^i = \alpha D_S + \beta D_H$. Finalement, la vidéo V est attribuée à la classe ayant la plus grande D_T^i.

2.3.3 Classification par un Modèle Bayésien de Régression Logistique

Selon R. Ksantini et Dubeau [71], que ce soit le taux d'erreur de classement ou le temps d'exécution pour la classification des images, la MBRL (Modèle Bayésien de Régression Logistique) performe mieux que plusieurs modèles existants (SVM, la RVM, etc.). Nous expérimentons alors ce modèle pour la classification des actions humaines. Ce choix est motivé par la disponibilité du code source.

La régression logistique classique ou bayésienne, a comme objectif la séparation par un hyperplan de deux groupes Ω_0 et Ω_1. Dans notre travail, nous considérons plus que deux groupes. Ceci nous pousse à utiliser le modèle de régression multinomial. Cette dernière consiste à comparer deux ensembles d'actions sous forme d'arbre d'actions. Par exemple, la base de données, décrite dans la section 2.4.1, représente six actions humaines (boxer, applaudir, agiter, marcher, courir, jogger). Nous divisons ces actions en deux groupes : des actions selon le mouvement des bras et d'autres selon celui des jambes. Ensuite, nous divisons les actions des bras en mouvement rapide (boxer) et mouvement lent (applaudir et agiter). Pour le deuxième groupe, concernant les actions selon les jambes, nous le divisons à son tour en groupes d'actions rapides (courir) avec celle plus lente (marcher et jogger). La figure 2.5 illustre notre modèle MBRL multinomial pour cette base de données.

Nous remarquons que cet exemple d'arbre d'actions (un contre un) nécessite en entrée une connaissance *a priori* et une bonne répartition des actions. Nous proposons alors de mettre plus l'accent sur chaque classe d'actions toute seule, en utilisant la stratégie (un contre tous). Pour cela, nous appliquons le MBRL binaire, avec un premier groupe contenant une classe donnée d'actions et l'autre groupe regroupant les classes d'actions restantes. Par exemple, pour la même base de données précédente, nous séparons d'abord

marcher de tout le reste (courir, jogger, boxer, applaudir et agiter) ensuite nous séparons courir du reste (marcher, jogger, boxer, applaudir et agiter), et ainsi de suite pour les autres classes d'actions. Pour un élément donné V, nous cherchons la classe d'actions la plus proche, qui est considérée comme sa classe d'appartenance. Dans nos résultats expérimentaux, nous remarquons que les deux stratégies décrites (un contre tous et un contre un) obtiennent des résultats similaires. Dans le cas où la stratégie un contre un est difficile à réaliser l'utilisation de notre stratégie (un contre tous), devient préférable.

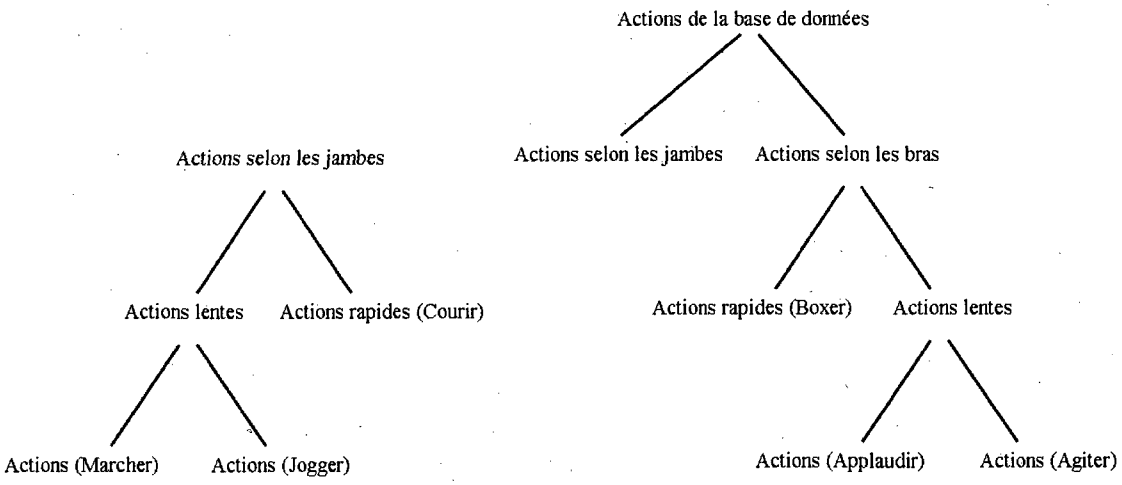


Figure 2.5 – Les catégories d'actions dans une vidéo.

Après avoir divisé nos deux groupes selon la stratégie (un contre tous), nous appliquons le MBRL binaire. Prenons par exemple, Ω_0 l'ensemble des données X_i^r qui provient de la classe marcher et Ω_1 l'ensemble des données X_j^{ir} des classes restantes (courir, applaudir, etc.) avec :

$$X_i^r = (\tilde{X}_{0,i}^r, X_{0,i}^r, \ldots, X_{J-1,i}^r, 1) \in \Omega_0 \tag{2.31}$$

$$X_j^{ir} = (\tilde{X}_{0,j}^{ir}, X_{0,j}^{ir}, \ldots, X_{J-1,j}^{ir}, 1) \in \Omega_1 \tag{2.32}$$

Pour la caractéristique PIST, J est l'ensemble de 128 données.

Les travaux de R. Ksantini et Dubeau [71] consistent à chercher une pseudométrique qui discrimine au mieux les deux groupes, et cela, par un hyperplan. Pour cette fin, il suffit de calculer les poids de cette pseudométrique $\tilde{\omega}$ et $\{\omega_k\}_{k=0}^{J-1}$. En utilisant la régression logistique bayésienne pour ce calcul, R. Ksantini et Dubeau [71] montrent que la pseudométrique est plus discriminative. De ce fait, une bonne séparation entre Ω_0 et Ω_1 est obtenue. Nous associons à la classe Ω_0, contenant n_0 premiers vecteurs propres PIST, une variable binaire $S_i^r = 0$. Pour les n_1 premiers vecteurs propres PIST contenu dans Ω_1, nous associons une variable binaire $S_j^{ir} = 1$. Avec ce modèle de régression logistique, les poids de la pseudométrique et un *intercept* v sont choisis pour maximiser la fonction logarithmique de vraisemblance.

$$\log(L(W = (\tilde{\omega_0}, \omega_0, \ldots, \omega_{J-1}, v))) = \sum_{i=1}^{n_0} \log(p_i^r) + \sum_{j=1}^{n_1} \log(p_j^{ir}) \tag{2.33}$$

avec la probabilité de pertinence p_i^r et de non-pertinence p_j^{ir} :

$$p_j^{ir} = P(S_j^{ir} = 1 | X_j^{ir}) = F(\tilde{\omega}_0 \tilde{X}_{0,j}^{ir} + \sum_{k=0}^{J-1} \omega_k X_{k,j}^{ir} + v) \tag{2.34}$$

$$p_i^r = P(S_i^r = 1 | X_i^r) = F(-\tilde{\omega}_0 \tilde{X}_{0,i}^r + \sum_{k=0}^{J-1} \omega_k X_{k,i}^r - v) \tag{2.35}$$

et avec une fonction logistique $F(x) = \frac{e^x}{1+e^x}$. Comme expliqué par R. Ksantini et Dubeau [71], il est difficile d'utiliser Fisher ou l'algorithme du gradient comme algorithme d'optimisation pour deux raisons, à savoir l'exponentielle dans la fonction de vraisemblance et la présence de zéros dans le vecteur de données. Comme Ω_0 et Ω_1 sont des ensembles de grande taille et de grandes dimensions, l'exponentielle tends vers 0, ce qui empêche la convergence. Afin de remédier aux deux premières limitations (exponentielle et zéro), R. Ksantini et Dubeau [71] lissent les paramètres à estimer, en admettant une distribution *a priori*. L'autre problème dû à la taille et la dimension des ensembles peut être résolu

en utilisant des transformations variationnelles qui simplifient le calcul des paramètres à estimer. Tout ceci a motivé le choix de la régression logistique bayésienne basée sur les transformations variationnelles (MBRL).

R. Ksantini et Dubeau [71] calculent la distribution *a posteriori* de W, avec $W = (\tilde{\omega_0}, \omega_0, \ldots, \omega_{J-1}, v)$ l'ensemble des paramètres des poids de la pseudométrique. Tel qu'en entrée, $W \approx \pi(W)$ avec π une distribution gaussienne *a priori*, de moyenne μ et de matrice de covariance Σ.

$$P(W|S_0 = 0, S_1 = 1) = \frac{\left[\sum_{x_0 \in \Omega_0, x_1 \in \Omega_1} \prod_{i=0}^{1} P(S_i = i|X_i = x_i, W) q_i(X_i = x_i)\right] \pi(W)}{P(S_0 = 0, S_1 = 1)} \tag{2.36}$$

avec

$$P(S_i = i|X_i = x_i, W) = F((2i-1)W^t x_i) \tag{2.37}$$

pour $i \in \{0, 1\}$.

En utilisant les estimations variationnelles et les inégalités de Jensen, la distribution *a posteriori* est estimée par

$$P(W|S_0 = 0, S_1 = 1) \geq \left[\prod_{i=0}^{1} F(\epsilon_i)\right] e^{\sum_{i=0}^{1}\left[\frac{E_{q_i}[H_i]-\epsilon_i}{2}\right] - \sum_{i=0}^{1}\left[\varphi(\epsilon_i)(E_{q_i}\left[H_i^2\right]-\epsilon_i^2)\right]} \tag{2.38}$$

avec E_{q_0} et E_{q_1} des espérances des distributions gaussiennes *a priori* q_0 et q_1 respectivement. $\varphi(\epsilon_i) = \frac{tanh(\frac{\epsilon_i}{2})}{4\epsilon_i}$ et $\{\epsilon_i\}_{i=0}^{1}$ sont les paramètres variationnels. De ce fait, la distribution *a posteriori* est estimée par une limite inférieure, et une distribution gaussienne de moyenne μ_{post} et de matrice de covariance Σ_{post}. Ces deux dernières sont estimées par une mise à jour de l'équation bayésienne :

$$(\Sigma_{post})^{-1} = (\Sigma)^{-1} + 2\sum_{i=0}^{1}\left[\varphi(\epsilon_i)(E_{q_i}\left[x_i(x_i)^t\right]\right] \tag{2.39}$$

$$\mu_{post} = \Sigma_{post}\left[(\Sigma)^{-1}\mu + \sum_{i=0}^{1}\left[(i-\frac{1}{2})(E_{q_i}[x_i]\right]\right] \tag{2.40}$$

La valeur de μ_{post} comporte les valeurs à estimer $\tilde{\omega}$, $\{\omega_k\}_{k=0}^{J-1}$ et *intercept* v. Pour trouver la classe d'appartenance d'une action x, il suffit de vérifier si $F((2i-1)W^t x_i)$ est supérieur à $0,5$.

De la même manière que la caractéristique PIST, nous appliquons le MBRL pour toute action représentée par la caractéristique CST. La différence majeure réside dans un prétraitement effectué sur les données, qui a pour but de représenter chaque vidéo par un seul vecteur de caractéristiques. Ce vecteur H_{moy} est la moyenne de tous les vecteurs du CST de la même vidéo.

2.3.4 Algorithme de classification par la MBRL

Notre algorithme a en entrée un ensemble T de vidéos d'apprentissage et une vidéo V, pour laquelle il faut déterminer la classe d'actions d'appartenance. En sortie, nous trouvons la classe d'actions de V. Avant d'effectuer la classification, nous passons par une étape d'apprentissage. Nous débutons par une extraction des caractéristiques, pour chaque élément de l'ensemble T, ainsi qu'une réduction des données. Nous calculons ensuite le vecteur moyen H_{moy} pour chaque vidéo de T. Les premiers vecteurs propres T_S des PIST de chaque vidéo de T sont rassemblés dans un groupe G_1. Tandis que le reste des vecteurs est mis dans un deuxième groupe G_2. Nous initialisons ensuite q_0, q_1 et la distribution *a priori* $\Pi(W)$. Par une méthode itérative, μ^S_{post} et Σ^S_{post} sont calculées pour la première caractéristique selon les équations (2.39) et (2.40). Nous rassemblons ensuite tous les vecteurs $T^i_{H_{moy}}$ d'une même action dans G_1, et le reste des vecteurs est placé dans un groupe G_2. Pour le CST, de même que les PIST, nous initialisons q_0, q_1 et la distribution *a priori* $\Pi(W)$, et nous calculons μ^H_{post} et Σ^S_{post}. Toute cette étape

d'apprentissage est refaite pour chaque classe d'actions (section 2.3.3).

Après l'apprentissage, nous obtenons deux vecteurs μ^S_{post} et μ^O_{post} pour chaque classe d'actions. Nous pouvons vérifier la classe d'appartenance d'un élément V (une vidéo d'une action humaine) par la classification. Comme pour chaque donnée en entrée, nous appliquons à V l'algorithme d'extraction des caractéristiques et celui de la réduction des données. Nous obtenons ainsi les deux vecteurs V_S et $V_{H_{moy}}$. Ensuite, nous calculons la probabilité de pertinence de V par rapport à la classe d'action marcher $p^{marcher}$, puis par rapport à la classe courir p^{courir}, ensuite jogger p^{jogger}, et ainsi de suite. Chaque probabilité de pertinence trouvée p^j, est composé de deux probabilités p^j_O et p^j_s. Ces deux dernières sont calculées selon l'équation 2.37. p^j_s est la probabilité de pertinence par rapport à la caractéristique PIST et p^j_O par rapport à la deuxième caractéristique CST tel que $p^j = \alpha p^j_O + \beta p^j_s$. α et β sont des paramètres de pondération trouvés expérimentalement. Enfin, nous attribuons V à la classe d'actions avec la plus grande probabilité p^j.

2.4 Résultats expérimentaux

Après, avoir présenté l'approche que nous effectuons pour la reconnaissance des actions humaines, nous testons le modèle décrit sur une des plus grandes bases de données existantes. Le but est de pouvoir estimer les différents paramètres du modèle, et d'évaluer ses performances dans la reconnaissance des actions. Les résultats obtenus lors de cette étape expérimentale sont indiqués et analysés dans cette partie. Pour déterminer l'apport de chaque caractéristique, nous évaluons les performances de plusieurs méthodes. Chaque méthode est la combinaison d'une caractéristique avec un modèle de classification. Dans la dernière section de cette partie, une comparaison, des performances de notre modèle avec celles des autres travaux, est dressée. Cela nous permet de relever les avantages de notre approche, ainsi que ses limitations par rapport aux travaux existants.

2.4.1 Les données

Notre modèle est validé en utilisant la plus grande base de données de vidéos d'actions humaines existante, dans laquelle chaque vidéo représente une action seule. Cette base de données, développée au sein du KTH par Schüldt *et al.* [77], se compose de six catégories d'actions humaines périodiques. Ces catégories représentent l'action de Marcher, de Jogger, de Courir, de Boxer, d'Agiter la main et d'Applaudir.

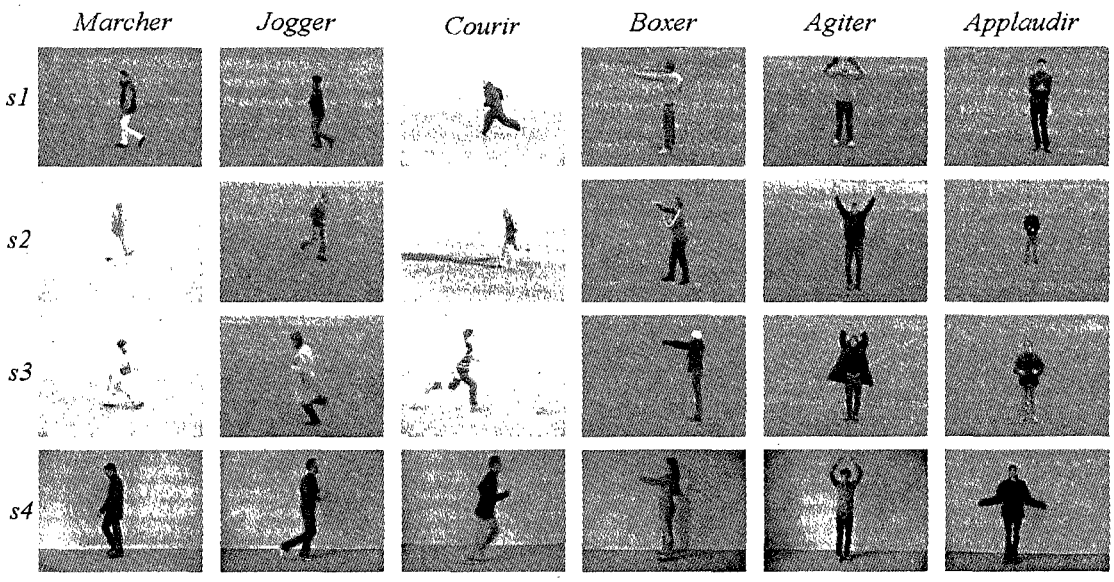


Figure 2.6 – Les différentes catégories d'actions selon les différents scénarios

Chaque action est jouée par 25 personnes différentes (hommes et femmes) selon 4 scénarios (figure 2.6). Le scénario s1 représente un environnement extérieur, le scénario s2 est dans un environnement extérieur avec des variations du zoom, le scénario s3 est aussi dans un environnement extérieur, mais où les sujets portent des habits différents tandis que le scénario s4 désigne un environnement intérieur. Donc, la base de données contient

600 vidéos, acquises avec une caméra fixe. Une vidéo contient un seul humain sur un fond homogène. Ces séquences vidéos sont de 25 images par secondes, à 256 niveaux de gris et une résolution de 160×120 pixels.

L'intérêt de choisir ces différents scénarios est celui de tester la robustesse d'un modèle de reconnaissance d'actions humaines périodiques, face à un changement d'échelle, d'environnement ou de personnes. Pour cette raison, plusieurs travaux [77, 60, 66, 45, 44, 29] ont choisi la base de données de Schüldt *et al.* pour tester et valider leur modèle de reconnaissance d'actions humaines.

De la même manière que les autres travaux [44, 60, 77], nous divisons aléatoirement la base de données en trois échantillons distincts. Un premier, pour effectuer l'apprentissage, un deuxième pour la validation et un troisième pour le test. Une telle décomposition est utile lorsque plusieurs méthodes sont testées et comparées. L'échantillon d'apprentissage permet d'abord de générer le modèle. Ce dernier est ensuite optimisé selon l'échantillon de validation. Finalement, l'erreur réelle de chaque modèle est évaluée à partir de l'échantillon de test.

Le premier échantillon, utilisé pour faire de l'**apprentissage**, est composé de vidéos de huit personnes effectuant les six types d'actions dans les quatre différents scénarios. Donc ce premier échantillon, qui est bien sûr choisi aléatoirement, est composé de 192 vidéos. Le but de cet échantillon est de représenter, tout en la distinguant, chaque catégorie d'actions. À partir des vidéos restantes, nous choisissons, et toujours aléatoirement, huit autres personnes pour représenter notre échantillon de **validation**. Cet échantillon est à son tour composé de 192 vidéos. Si des paramètres interviennent dans le modèle, alors ils sont estimés au cours de cette étape. L'échantillon de **test** est donc composé du reste de la base de données, c'est-à-dire 216 vidéos. Ceci représente 40% de la base de Schüldt *et al.*. C'est à partir de ce dernier échantillon que nous mesurons les performances du modèle.

2.4.2 La méthodologie

Pour évaluer les performances de notre modèle, nous testons différentes méthodes sur la base de données de Schüldt *et al.*. L'objectif est de mettre en valeur l'apport de chaque caractéristique extraite ainsi que les performances des différents modèles de classification. Dans cette section, nous décrivons d'abord les méthodes utilisées lors de notre expérimentation, ensuite nous indiquons, pour certaines de ces dernières, les principaux paramètres à estimer.

Pour représenter une action humaine dans une vidéo, le choix de la caractéristique est très important. Pour déterminer les performances de cette caractéristique choisie pour notre modèle, nous testons deux caractéristiques (section 2.1) : le PIST (Points d'intérêts Spatiao-Temporles) et le CSST (Contour et SIFT Spatio-Temporel).

Une fois les caractéristiques extraites d'une vidéo, l'action est affectée à sa catégorie, selon un modèle de classification. Comme expliqué dans la section 2.3, nous optons pour les modèles de classifications suivants : le MBRL (Modèle Bayésien de Régression Logistique), le Kppvc (distance cosinus) et le Kppve (distance euclidienne).

Avec ces caractéristiques et ces modèles de classification, nous obtenons six méthodes différentes pour faire de la reconnaissance d'actions humaines. Pour la plupart de ces méthodes, certains paramètres doivent être déterminés, afin d'avoir des résultats optimaux. Pour nos tests, nous avons deux paramètres principaux, le nombre de voisins k pour le modèle Kppv, ainsi que la valeur du poids α de chaque caractéristique dans le cas d'une combinaison de deux caractéristiques (CSST).

- Le **nombre** k est le premier paramètre à estimer. Lorsque k prend 1 comme valeur, le temps de calcul diminue, mais le taux d'erreur de classification augmente, et vice-versa. Donc, l'objectif est de déterminer le nombre k des plus proches voisins qui offre un compromis acceptable.

- La valeur du **poids** α est le deuxième paramètre à estimer. Le CSST implique un poids α à attribuer à la contribution de chaque caractéristique. La valeur de α est comprise entre 0 et 1. Lorsque cette dernière prend 1 comme valeur, la classification d'une vidéo se fait seulement grâce à la première caractéristique, et c'est le contraire si $\alpha = 0$.

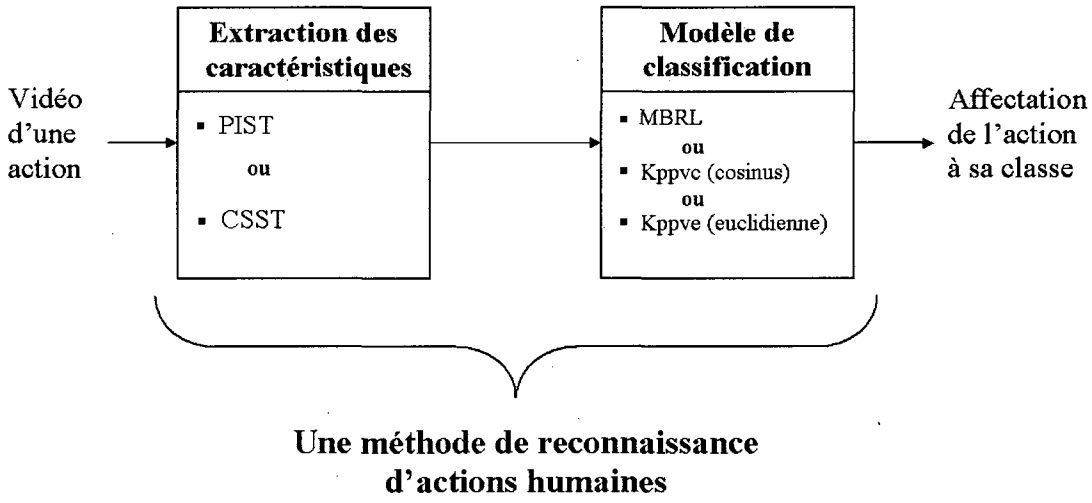


Figure 2.7 – Les différentes méthodes pour le test.

Un schéma récapitulatif de cette section est illustré par la figure 2.7. Ce schéma illustre les différentes méthodes utilisées lors de nos tests expérimentaux. À partir des résultats obtenus, nous évaluons la performance de chaque méthode à partir du taux de bien classé de chaque catégorie d'actions, ainsi qu'une moyenne de ces taux. Cette dernière est considérée comme la mesure de performance principale. Dans la section suivante, nous présentons les résultats des expérimentations de ces six méthodes.

2.4.3 Les expérimentations

Lors de notre expérimentation, les six méthodes sont testées. Nous choisissons de présenter ces tests selon chacune des deux caractéristiques. La 1re expérimentation décrit les résultats obtenus selon la caractéristique PIST, alors que la 2e expérimentation regroupe ceux de la combinaison CSST. Ainsi, nous évaluons les performances de chaque caractéristique selon les différents modèles de classifications. Donc, chaque expérimentation regroupe les résultats de trois méthodes. Pour comparer les performances de chaque méthode, nous considérons un tableau contenant le taux de bien classé pour chaque catégorie de la base de données, ainsi qu'une moyenne de ces taux. À la fin de cette section, nous présentons plus en détail la méthode du CSST avec le MBRL qui obtient les meilleures performances dans la reconnaissance d'actions humaines.

1re expérimentation

Cette expérimentation décrit les résultats de la caractéristique PIST selon les modèles de classifications MBRL, Kppvc et Kppve.

Les résultats de la reconnaissance d'actions humaines, selon le modèle de MBRL, sont exposés dans le tableau 2.2. Ce dernier contient les taux de bien classé pour chaque groupe, ainsi qu'une moyenne de ces taux. Nous obtenons un taux moyen de bien classé de $35,8\%$. Ces résultats montrent que la caractéristique PIST seule n'arrive pas à classer correctement les actions humaines selon le modèle de MBRL. À part la catégorie Courir qui atteint $69,4\%$, le modèle ne classe pas plus de quatre vidéos sur dix en moyenne. Ces résultats sont dû principalement à l'utilisation de l'ACP où ce dernier élimine les données récurrentes et d'où les PIST qui se répètent d'une image à une autre de la même vidéo. Alors que ces points sont très importants pour la reconnaissance de l'action.

Les résultats de la classification d'actions humaines selon le modèle de Kppvc sont illustrés

Marcher	Jogger	Courir	Boxer	Applaudir	Agiter	**Moyenne**
36,1%	25%	69,4%	33,4%	25,7%	25%	**35,8%**

Tableau 2.2 – Les taux de bien classé pour la caractéristique PIST avec le MBRL.

par la figure 2.8. Nous obtenons un taux moyen de bien classé de $34,1\%$ pour $k = 1$, de $37,4\%$ pour $k = 8$ et de $38,6\%$ pour $k = 16$. Le modèle Kppvc classe pas plus de quatre vidéos sur dix en général. En fixant, le nombre de voisins $k = 16$, nous obtenons le plus haut taux de bien classé. Le tableau 2.3 reprend ces taux de bien classé pour chaque catégorie, ainsi qu'une moyenne de ces taux. D'après ce tableau, nous remarquons qu'avec la caractéristique PIST seule, le modèle Kppvc classe à peu près sept vidéos sur dix de la catégorie Jogger. Pour les autres catégories, il ne dépasse pas le taux de 40% de bien classé.

Marcher	Jogger	Courir	Boxer	Applaudir	Agiter	**Moyenne**
25%	69,4%	36,1%	25%	37,1%	38,9%	**38,6%**

Tableau 2.3 – Les taux de bien classé pour la caractéristique PIST avec le modèle Kppvc, $k = 16$.

Les résultats de la classification d'actions humaines selon le modèle de Kppve sont illustrés par la figure 2.9. Nous obtenons un taux moyen de bien classé de $36,2\%$ pour $k = 1$, de $28,7\%$ pour $k = 8$ et de $33,9\%$ pour $k = 16$. Cette fois-ci, nous remarquons que le Kppve donne un meilleur taux de bien classé pour un nombre de voisins $k = 1$. Les résultats de cette classification sont illustrés par le tableau 2.4. Quelle que soit la catégorie, le Kppve ($k = 1$) n'arrive pas à classer correctement plus que cinq vidéos sur dix.

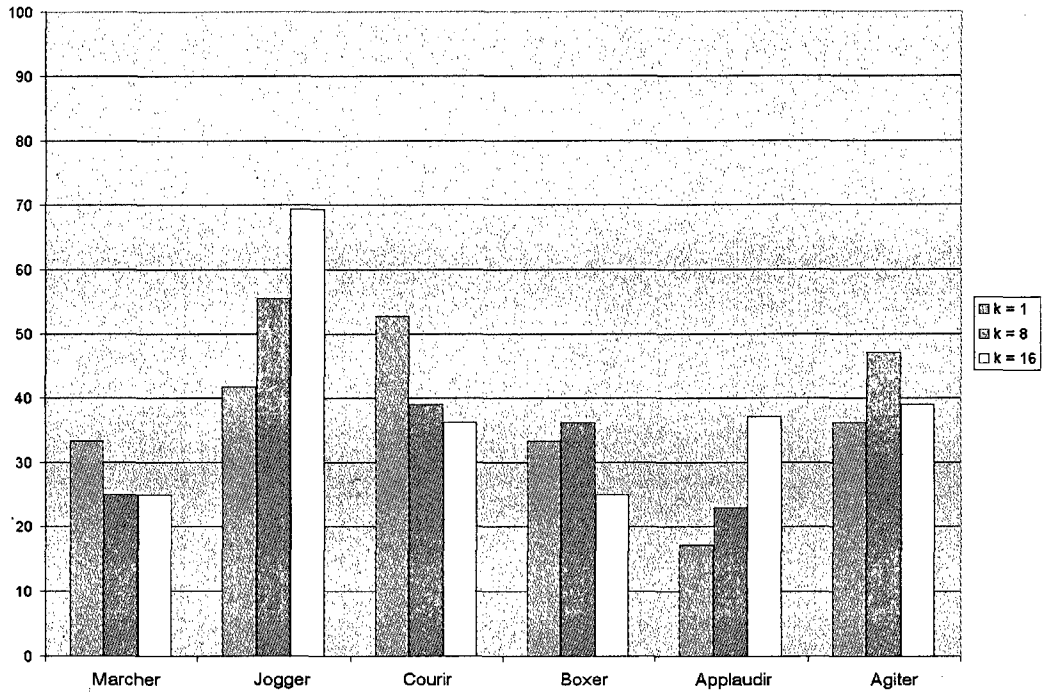


Figure 2.8 – Les taux de bien classé pour la caractéristique PIST avec le modèle Kppvc, selon différentes valeurs du nombre k.

Marcher	Jogger	Courir	Boxer	Applaudir	Agiter	**Moyenne**
41,7%	30,6%	44,4%	41,7%	14,3%	44,4%	**36,2%**

Tableau 2.4 – Les taux de bien classé pour la caractéristique PIST avec le modèle Kppve, $k = 1$.

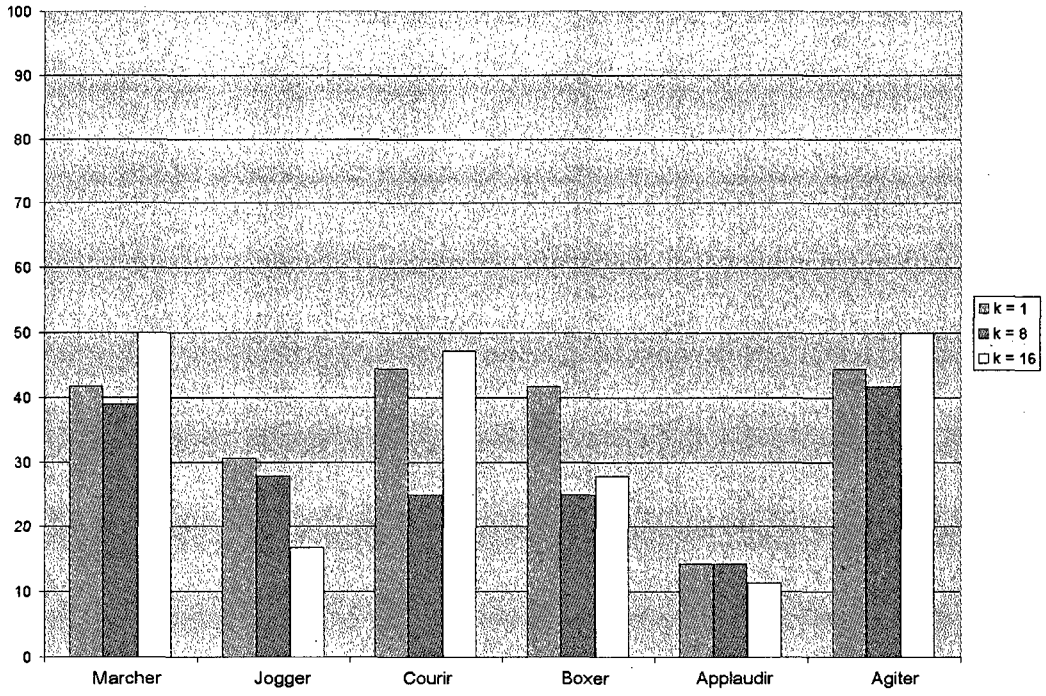


Figure 2.9 – Les taux de bien classé pour la caractéristique PIST avec le modèle Kppve, selon différentes valeurs du nombre k.

Comme conclusion pour cette première expérimentation, nous estimons que la caractéristique PIST conduit à une classification ne dépassant pas quatre vidéos sur dix, toutes les catégories confondues. Les résultats obtenus montrent que cette caractéristique performe mieux avec le Kppvc avec un $k = 16$. Malgré cela, le taux obtenu, qui est de $38,6\%$, reste insuffisant.

2^{e} expérimentation

Cette seconde expérimentation décrit les résultats pour la nouvelle caractéristique. Cette dernière est la combinaison CSST (Contour et Sift Spatio-Temporel). De la même manière que l'expérimentation précédente, nous présentons les résultats selon les modèles de classifications MBRL, Kppvc et Kppve. Sachant que le CSST résulte d'une combinaison de caractéristiques, l'estimation du paramètre α (section 2.4.2) devient nécessaire. Ce paramètre, qui représente le poids de chaque caractéristique, est estimé pendant l'étape de validation. Pendant cette étape, nous faisons varier le paramètre α de 0 à 1, avec un pas de $0,01$. Pour chacune de ces valeurs, nous calculons le taux de bien classé pour chaque catégorie de l'échantillon de validation. Nous retenons finalement la valeur α qui offre le meilleur compromis entre le taux moyen de bien classé et l'écart type moyen.

Les résultats de la classification d'actions humaines, selon le MBRL, sont exposés dans cette partie. Le tableau 2.5 contient les taux de bien classé pour chaque groupe, ainsi qu'une moyenne de ces taux. Notons que nous obtenons ces résultats pour un $\alpha = 0.76$. Cette combinaison CSST donne de bons résultats avec le MBRL. Le taux moyen de bien classé obtenu est de $81,4\%$ pour la totalité d'actions humaines. Nous remarquons aussi que la classification des actions de la catégorie Jogger est presque parfaite (un taux de $97,2\%$).

Les résultats de la classification d'actions humaines selon le modèle de Kppvc sont illustrés par la figure 2.10. Lors de l'étape de validation, nous obtenons les meilleurs résultats pour

Marcher	Jogger	Courir	Boxer	Applaudir	Agiter	**Moyenne**
77,8%	97,2%	80,6%	80,6%	74,3%	77,8%	**81,4%**

Tableau 2.5 – Les taux de bien classé pour la caractéristique CSST selon le MBRL, $\alpha = 0.76$.

un $\alpha = 0.61$. Nous constatons un taux moyen de bien classé de $80,4\%$ pour $k = 1$, de $83,5\%$ pour $k = 8$ et de $77,6\%$ pour $k = 16$. Le modèle Kppvc classe en moyenne huit vidéos sur dix en général. En fixant, le nombre de voisins $k = 8$, nous obtenons le plus haut taux de bien classé. Le tableau 2.6 reprend ces taux de bien classé pour chaque catégorie, ainsi qu'une moyenne de ces taux. Les résultats de cette méthode sont acceptables, ce qui est confirmé par le taux de bien classé moyen de $83,5\%$. En fixant $\alpha = 0,61$ et le nombre de voisins $k = 8$, nous reconnaissons toutes les actions de Courir, Marcher, Jogger, et Agiter.

Marcher	Jogger	Courir	Boxer	Applaudir	Agiter	**Moyenne**
94,5%	97,2%	100%	75%	37,1%	97,2%	**83,5%**

Tableau 2.6 – Les taux de bien classé pour la caractéristique CSST selon le modèle Kppvc avec $\alpha = 0.61$ et $k = 8$.

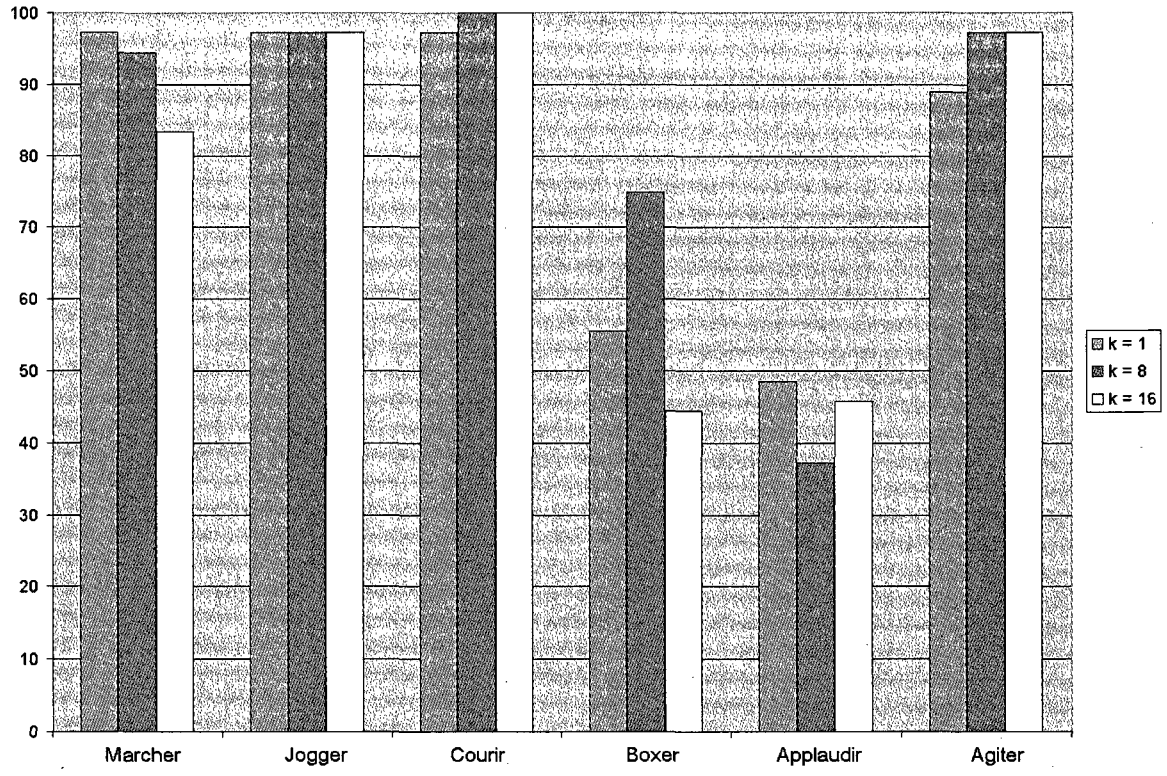


Figure 2.10 – Les taux de bien classé pour la caractéristique CSST le modèle Kppvc selon différentes valeurs du nombre de voisins k, $\alpha = 0.61$.

Les résultats de la classification d'actions humaines selon le modèle de Kppve sont illustrés par la figure 2.11. Ces résultats sont calculés pour un $\alpha = 0.61$. Nous obtenons un taux moyen de bien classé de $49,1\%$ pour $k = 1$, de $42,7\%$ pour $k = 8$ et de $41,3\%$ pour $k = 16$. Cette fois-ci, nous remarquons que le modèle Kppve donne un meilleur taux de bien classé pour un nombre de voisins $k = 1$. Les résultats de cette classification sont illustrés par le tableau 2.7. Pour cette méthode, le meilleur taux moyen de bien classé est trouvé en considérant un seul voisin ($k = 1$). En général, ce modèle classe une vidéo sur deux, dans sa catégorie correspondante.

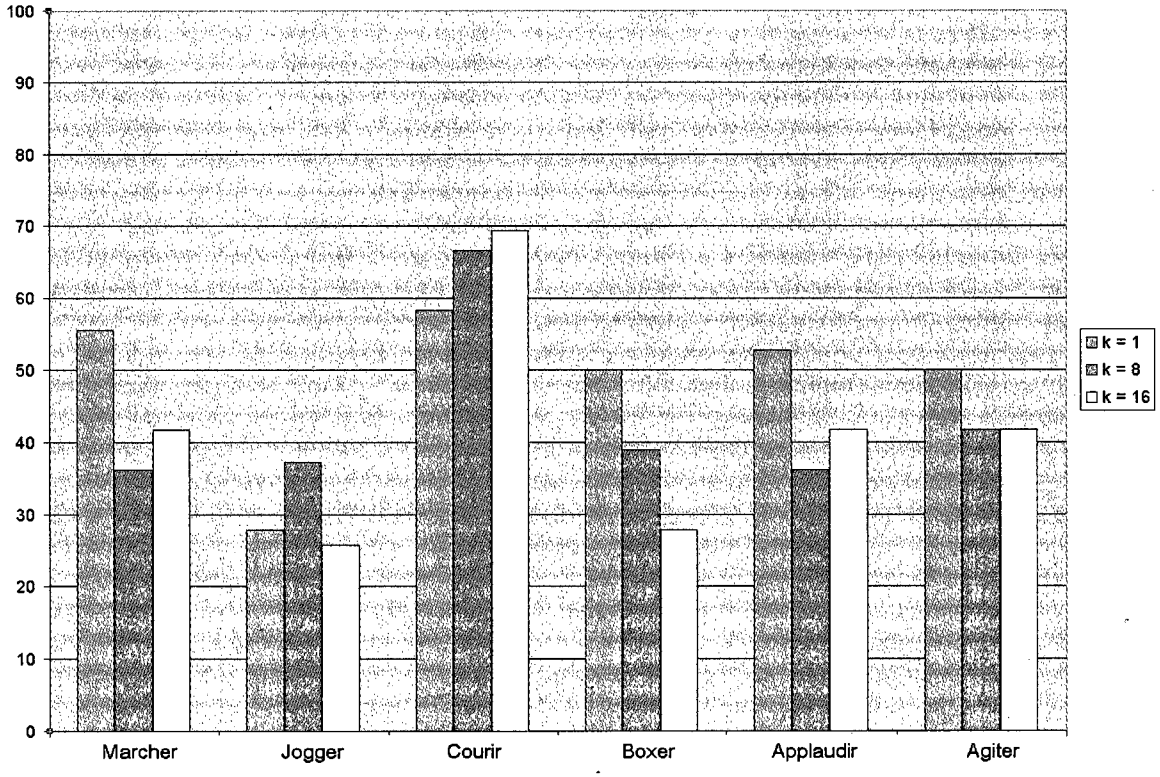


Figure 2.11 – Les taux de bien classé pour la caractéristique CSST selon le modèle Kppve selon différentes valeurs du nombre k, $\alpha = 0.61$.

Marcher	Jogger	Courir	Boxer	Applaudir	Agiter	**Moyenne**
55,5%	27,8%	58,3%	50%	52,8%	50%	**49,1%**

Tableau 2.7 – Les taux de bien classé pour la caractéristique CSST selon le modèle Kppve avec $\alpha = 0.61$ et $k = 1$.

D'après les résultats obtenus, nous constatons que la combinaison CSST avec le Kppvc performe mieux qu'avec le Kppve, quelque soit le nombre de voisins k. Pour $k = 1$, les performances ne sont pas influencées seulement par le choix de la distance. En effet, pour le PIST avec le Kppv, nous obtenons un taux moyen de bien classé de $36,2\%$ pour la distance euclidienne, et de $34,1\%$ pour la distance cosinus. Alors que le taux moyen de bien classé, pour la CSST, est de $49,1\%$ pour la distance euclidienne, et de $80,4\%$ pour la distance cosinus. Comme conclusion sur l'utilisation du Kppv, il est certain que la distance cosinus est meilleure que la distance euclidienne, surtout pour un nombre de voisins élevé. Alors qu'il est préférable de chercher la meilleure distance entre les deux lors de l'étape de la validation, lorsque le nombre de voisins est faible (base de données réduite). Ces conclusions confirment le travail de Qian *et al.* [70].

Avec la caractéristique CSST, nous constatons que le taux moyen de bien classé s'est amélioré, quel que soit le modèle de classification utilisé. En effet, les résultats obtenus par les méthodes précédentes confirment l'apport de cette nouvelle caractéristique. Nous remarquons d'abord que le modèle de classification Kppv classe mieux les actions humaines, lorsqu'il est utilisé avec une distance cosinus. La combinaison CSST est idéale pour la reconnaissance de vidéos d'actions humaines, surtout avec notre modèle MBRL. Ce dernier se démarque du modèle Kppvc, avec un nombre de voisins $k = 8$, par ses résultats cohérents par rapport aux catégories d'actions. Même si le modèle Kppvc a un taux moyen de $83,5\%$ alors que le MBRL obtient un taux de $81,4\%$, nous choisissons d'utiliser ce dernier pour sa stabilité et pour sa rapidité d'exécution.

Comme mentionné dans la section 2.3.3 nous avons expérimenté aussi le MBRL selon la stratégie arbre d'actions. Nous obtenons des résultats similaires à la MBRL selon la stratégie un contre tous. Ceci est confirmé par le tableau 2.8. Nous obtenons une moyenne de $79,6\%$, tandis que nous obtenons un taux moyen de $81,4\%$ par la méthode MBRL un contre tous. Les résultats sont invariants aux catégories d'actions, et cela, pour les deux méthodes. Ces résultats sont presque similaires avec un meilleur taux pour la MBRL (un contre tous).

Marcher	Jogger	Courir	Boxer	Applaudir	Agiter	**Moyenne**
76,4%	96,3%	78,6%	81,1%	63,7%	72,5%	**79,6%**

Tableau 2.8 – Les taux de bien classé pour la caractéristique CSST selon le MBRL (arbre de décision), $\alpha = 0.76$.

Nous optons alors pour le CSST avec le MBRL et la stratégie un contre tous comme méthode idéale pour faire de la reconnaissance d'actions humaines. Dans la section qui suit, nous présentons les résultats détaillés de cette méthode.

La méthode du CSST avec le MBRL

Dans ce qui suit, nous détaillons les résultats de la méthode du CSST avec la MBRL selon la stratégie un contre tous. L'objectif est de mettre en valeur les avantages et les inconvénients d'une telle méthode. Pour cette raison, nous commençons par présenter la matrice de confusion pour cette méthode. Une telle mesure indique clairement les taux de bien classé pour chaque catégorie, et aussi les taux de confusion de mal classé. Nous poursuivons par une présentation des résultats selon chaque scénario isolé, ce qui permet d'évaluer les performances de notre modèle selon la nature des vidéos prises. À la fin, nous présentons les résultats du modèle selon la classification de chaque catégorie d'actions humaines.

La matrice de confusion

Comme mesure d'évaluation, la matrice de confusion est souvent utilisée pour l'évaluation des performances d'un modèle de classification. Le tableau 2.9 représente la matrice de confusion du modèle de classification MBRL, avec comme caractéristique le CSST. Nous effectuons toujours une classification selon le même échantillon de test.

	Marcher	Jogger	Courir	Boxer	Applaudir	Agiter
Marcher	**77,8%**	5,5%	2,8%	5,5%	0%	8,4%
Jogger	0%	**97,2%**	0%	2,8%	0%	0%
Courir	0%	16,6%	**80,6%**	0%	2,8%	0%
Boxer	5,5%	0%	0%	**80,6%**	8,4%	5,5%
Applaudir	8,6%	2,8%	0%	8,6%	**74,3%**	5,7%
Agiter	2,8%	0%	0%	0%	19,4%	**77,8%**

Tableau 2.9 – La matrice de confusion selon la MBRL, avec la CSST.

Comme nous l'avons déjà mentionné avant nous obtenons un taux de classification moyen de 81.4%. Selon la matrice de confusion, 16.6% des vidéos représentant des personnes qui courent (Courir) sont attribués à la catégorie Jogger. Cela peut s'expliquer par le fait que les deux mouvements sont presque similaires. Même chose pour la catégorie Agiter, où 19.4% des vidéos à classer ont été confondues avec celles de la catégorie Applaudir.

Pour vérifier ces confusions, et vu la nature des actions de notre base de données, nous regroupons ces catégories en deux classes distinctes. La première représente les actions avec un mouvement de jambes (Marcher, Jogger et Courir). Et la seconde représente les actions selon un mouvement du bras. Le tableau 2.10 montre que la confusion est très petite dans ce cas, elle est en moyenne de 6.5% pour chacune des deux classes. Nous pouvons en déduire que notre modèle distingue parfaitement les actions humaines où le corps en entier est en mouvement, des actions qui nécessitent le mouvement des bras seulement.

	Bien classé	Mal classé
Classe 1	93,5%	6,5%
Classe 2	6,6%	93,4%

Tableau 2.10 – Les taux de bien classé pour le regroupement des catégories d'actions en deux classes.

Les résultats par scénarios

Comme la base de données est réalisée selon quatre scénarios différents nous présentons les performances de notre modèle selon chacun de ces scénarios. Avec la même mesure d'évaluation, nous comparons les performances de classification selon chaque scénario. Les scénarios sont définis dans la section 2.4.1 de ce chapitre. Chacun représente le quart de la base de données utilisée.

Le premier scénario, appelé s1, représente une action effectuée à l'extérieur. Le tableau 2.11 reprend la matrice de confusion pour ce scénario. Pour cette classification, l'échantillon utilisé comprend seulement des vidéos réalisées selon le scénario s1. Nous obtenons de très bons résultats, puisque le taux moyen de bien classé, toute catégorie confondue, est de $90,7\%$. Un tel résultat montre que notre modèle classe parfaitement des actions réalisées à l'extérieur, surtout les actions de Boxer et de Jogger, où le taux moyen pour chacune est de 100%.

	Marcher	Jogger	Courir	Boxer	Applaudir	Agiter
Marcher	**77,8%**	11,1%	0%	0%	0%	11,1%
Jogger	0%	**100%**	0%	0%	0%	0%
Courir	0%	0%	**88,9%**	0%	11,1%	0%
Boxer	0%	0%	0%	**100%**	0%	0%
Applaudir	0%	0%	0%	0%	**88,9%**	11,1%
Agiter	0%	0%	0%	0%	11,1%	**88,9%**

Tableau 2.11 – La matrice de confusion selon le scénario s1, pour la MBRL avec la CSST.

En effectuant une variation du zoom sur les personnes qui effectuent les actions, le scénario s2 est le plus complexe des scénarios utilisés dans la base de données. Nous réalisons une classification des vidéos de ce scénario, qui sont estimées à 54 vidéos. Le résultat de cette classification est représenté par la matrice de confusion, illustrée par le tableau 2.12. En effet, la complexité de ce scénario est démontrée par les résultats obtenus. Le taux moyen de bien classé est de $74,1\%$ pour les vidéos réalisées sous ce scénario. Ce taux est plus bas que la moyenne que nous avons obtenue pour la totalité de la base de données. Nous remarquons aussi que les vidéos appartenant à la catégorie Agiter sont mal classées $(44,5\%)$, et généralement confondues $(44,4\%)$ avec la catégorie Applaudir. Cela s'explique par le fait que ces actions sont similaires, surtout avec un effet de zoom variant.

	Marcher	Jogger	Courir	Boxer	Applaudir	Agiter
Marcher	**88,9%**	0%	0%	0%	0%	11,1%
Jogger	0%	**88,9%**	0%	11,1%	0%	0%
Courir	0%	22,2%	**77,8%**	0%	0%	0%
Boxer	11,1%	0%	0%	**66,7%**	11,1%	11,1%
Applaudir	0%	0%	0%	11,1%	**77,8%**	11,1%
Agiter	11,1%	0%	0%	0%	44,4%	**44,5%**

Tableau 2.12 – La matrice de confusion selon le scénario s2, pour la MBRL avec la CSST.

Le troisième scénario s3 est à son tour réalisé à l'extérieur. La différence avec le scénario s1 est le changement d'apparence des personnes effectuant les actions. Ce changement est dû particulièrement aux changements de vêtements. La classification de ses vidéos selon notre modèle est exprimée dans le tableau 2.13. Cette fois-ci, nous obtenons un taux moyen de bien classé de $75,5\%$. À part, la catégorie Jogger, qui a un taux de 100% de bien classé, et la catégorie Boxer, avec un taux de $88,9\%$, le modèle ne distingue pas efficacement les autres catégories.

	Marcher	Jogger	Courir	Boxer	Applaudir	Agiter
Marcher	**55,6%**	11,1%	0%	22,2%	0%	11,1%
Jogger	0%	**100%**	0%	0%	0%	0%
Courir	0%	22,2%	**77,8%**	0%	0%	0%
Boxer	0%	0%	0%	**88,9%**	0%	11,1%
Applaudir	25%	12,5%	0%	12,5%	**50%**	0%
Agiter	0%	0%	0%	0%	22,2%	**77,8%**

Tableau 2.13 – La matrice de confusion selon le scénario s3, pour la MBRL avec la CSST.

Le dernier scénario s4 se distingue des autres par son environnement. En effet, les vidéos de ce scénario sont les seuls à être réalisées dans un environnement intérieur. Le tableau 2.14 reprend les résultats de la classification de cette catégorie de vidéos. Le taux moyen de classification est de $85,2\%$, ce qui prouve que notre modèle s'adapte parfaitement à un environnement intérieur. Certaines catégories, comme Jogger et Boxer, obtiennent des taux de classification parfaite, avec un taux moyen de 100%. Les actions appartenant à la catégorie Courir sont soit bien classées avec un taux de $77,8\%$ ou sinon confondues avec la catégorie Jogger avec un taux de $22,2\%$.

	Marcher	Jogger	Courir	Boxer	Applaudir	Agiter
Marcher	**88,9%**	0%	11,1%	0%	0%	0%
Jogger	0%	**100%**	0%	0%	0%	0%
Courir	0%	22,2%	**77,8%**	0%	0%	0%
Boxer	11,1%	0%	0%	**66,7%**	22,2%	0%
Applaudir	11,1%	0%	0%	11,1%	**77,8%**	0%
Agiter	0%	0%	0%	0%	0%	**100%**

Tableau 2.14 – La matrice de confusion selon le scénario s4, pour la MBRL avec la CSST.

Les résultats selon les catégories d'actions

Dans ce qui suit, nous présentons le taux de bien classé de notre modèle en fonction du la catégorie de l'action humaine de la base de données de Schüldt *et al.*. À l'aide de taux de bien classé ou de confusion, nous comparons les résultats obtenus pour les différentes catégories. Comme pour les résultats précédents, ces taux sont calculés à partir de l'échantillon de test (192 vidéos). Le tableau 2.15 illustre le taux de confusion pour chaque catégorie d'actions. Le taux de confusion représente les vidéos qui contiennent des actions différentes de la catégorie d'actions, où elles sont classées. Par exemple, seulement $3,4\%$ des actions, ne contenant pas de personne qui boxe, sont classés dans la catégorie Boxer.

	Classée (Faux)	Non Classée (Vrai)
Marcher	3,4%	96,6%
Jogger	5%	95%
Courir	0,5%	99,5%
Boxer	3,4%	96,6%
Applaudir	6,1%	93,9%
Agiter	6,2%	93,8%
Moyenne	**4,1%**	**95,9%**

Tableau 2.15 – Les taux de confusion par catégorie d'actions.

Pour des vidéos qui appartiennent à la catégorie **Marcher**, en moyenne, huit vidéos sur dix sont reconnues par notre modèle, tout scénario confondu. Alors que pour les autres vidéos, qui n'appartiennent pas à cette catégorie, le taux de confusion est de $3,4\%$ (tableau 2.15). Donc, le modèle distingue correctement cette catégorie d'actions. Nous remarquons aussi qu'une action représentant une personne qui marche est distinguée plus souvent dans un environnement intérieur $(88,9\%)$.

La catégorie **Jogger** prête à confusion avec les catégories de Marcher et de Courir. Considérée comme une action intermédiaire, entre Courir et Marcher, cette catégorie

d'action est la mieux classée par notre modèle. Son taux moyen de bien classé est de 97,2% (tableau 2.9), valable pour l'échantillon au complet. Les vidéos d'action de Jogger sont reconnues à 100% dans tous les scénarios, et à peu près neuf vidéos sur dix sont reconnues lorsqu'il y a une variation de zoom dans la vidéo. Avec d'aussi bons résultats, cette catégorie obtient un taux de 5,1% (tableau 2.15) de confusion par rapport aux vidéos n'appartenant pas à la catégorie Jogger.

Le taux de bien classé moyen pour la catégorie **Courir** est de 80,6% (tableau 2.9), ce qui est considéré bon pour des actions aussi rapides. L'environnement n'influence pas vraiment la classification de ce genre d'actions par notre modèle, qui reste assez stable. Selon le tableau 2.15, le taux de confusion est de 0.5% pour une vidéo qui n'appartient pas à cette catégorie. Ce taux est le plus bas pour toutes les catégories, ce qui montre que presque aucune autre action ne peut être considérée comme une action de Courir.

La catégorie **Boxer** présente un taux élevé de bien classé, qui est de 80,6% en moyenne (tableau 2.9). Les actions de cette catégorie sont reconnues à 100% lorsque celles-ci sont effectuées dans un environnement extérieur. Ce taux baisse à 66,7% lorsqu'il y a un effet de zoom dans la vidéo de l'action, ou lorsque celle-ci se trouve dans un environnement intérieur. Pour la confusion, seulement 3,4% des actions, n'appartenant pas à Boxer, sont considérées comme des actions représentant une personne qui boxe (tableau 2.15).

Même si son taux de bien classé est le plus bas, en comparaison avec les autres catégories, la catégorie **Applaudir** classe correctement plus de 74% (tableau 2.9) des vidéos de Applaudir. Comme la plupart des autres catégories, elle a un taux de 88,9% de bien classé pour des vidéos du scénario s1. Nous remarquons aussi que le tableau 2.15 indique un taux de confusion de 6,1%. Ce dernier s'explique par le fait que plusieurs vidéos, appartenant à la catégorie Agiter, ont été assimilées à la catégorie Applaudir.

Pour notre modèle de classification d'actions humaines, la catégorie **Agiter** obtient un taux de bien classé de $77,8\%$ en moyenne (tableau 2.9). Comme la catégorie Jogger, le modèle arrive à reconnaitre toutes les vidéos de cette catégorie, lorsqu'elles sont réalisées à l'intérieur. Nous remarquons, selon le tableau 2.15, que son taux de confusion est de $6,2\%$, pour des vidéos n'appartenant pas à la catégorie Agiter.

2.4.4 Comparaison avec les autres travaux

Plusieurs travaux se sont intéressés à la reconnaissance d'actions humaines. Avec des caractéristiques et des modèles de classifications différents, ces travaux ont testé leur performance de reconnaissance. Pour cela, le choix d'une base de données complète et variée est nécessaire. Ces dernières années, la plupart des travaux ont utilisé la base de données KTH développée par Schüldt *et al.* (section 2.4.1). Dans cette section, nous situons notre travail par rapport aux différents travaux existants. Pour ce but, nous formons un tableau regroupant les taux de bien classé pour chaque catégorie, ainsi qu'une moyenne de ces taux.

Pour la suite, nous avons choisi sept travaux différents de reconnaissance d'actions humaines. Pour déterminer les performances de leur modèle, les chercheurs ont utilisé plusieurs classificateurs. Dans notre comparaison, nous considérons le meilleur résultat obtenu dans chacun de ces travaux. La plupart optent pour des modèles de classifications connus (SVM, Kppv, etc.). Cependant, certains chercheurs développent leur propre modèle, comme le cas de Yeo *et al.* qui utilise le NZMS [1]. Le tableau 2.16 présente ces sept travaux, avec leurs caractéristiques choisies et leur modèle de classification utilisé. En effectuant des tests sur la totalité de la base de données, certains travaux ont considéré tous les scénarios de la base, même les plus complexes. Cependant, quelques travaux se

1. NZMS (Non-Zero Motion Block Similarity) est une mesure développée par Yeo *et al.*[87], qui permet d'éliminer la similarité entre les régions non significatives.

sont contentés d'une partie de la base de données. Ce choix de la complexité de la base est précisé dans le tableau 2.16, ainsi que le taux moyen de bien classé pour chaque travail. Nous appelons notre méthode basée sur le CSST avec le MBRL notre "méthode 1", alors que la "méthode 2" désigne la méthode basée sur le CSST avec le Kppvc.

	Caractéristique	Classificateur	Base de données	Taux
Ke *et al.* [44]	Volume 3D spatio-temporel	Cascade de filtres	Base du KTH avec complexité	**62,9%**
Schüldt *et al.* [77]	Points intérêts spatio-temporels	SVM	Base du KTH avec complexité	**71,7%**
Meng *et al.* [60]	Histogramme du MHI	SVM	Base du KTH avec complexité	**80,3%**
Dollar *et al.* [29]	Points intérêts spatio-temporels	SVM	Base du KTH sans complexité	**81,1%**
Notre méthode 1	CSST	MBRL	Base du KTH avec complexité	**81,4%**
Niebles *et al.* [66]	Mots spatio-temporels	pLSA	Base du KTH avec complexité	**81,5%**
Kienzle *et al.* [45]	Points intérêts spatio-temporels	SVM	Base du KTH avec complexité	**82,8%**
Notre méthode 2	CSST	Kppv	Base du KTH avec complexité	**83,5%**
Yeo *et al.* [87]	Vecteur mouvement avec NZMS	Kppv	Base du KTH sans complexité	**86%**

Tableau 2.16 – Comparaison des modèles de reconnaissance d'actions humaines selon différents travaux.

Nous constatons que nos méthodes, basées sur le CSST, donnent de bons résultats. En considérant la base de données de Schüldt *et al.* avec sa complexité, nous remarquons que les caractéristiques choisies, dans notre travail, sont performantes dans la reconnaissance d'actions humaines. Par rapport aux autres travaux, les résultats de nos deux méthodes peuvent être considérés comme les meilleurs.

Pour évaluer en détail les performances de nos caractéristiques, nous comparons les taux de bien classé pour chaque catégorie de la base de données. Cette comparaison, illustrée par le tableau 2.17, est effectuée par rapport aux travaux qui ont considéré la base du KTH avec sa complexité.

	Marcher	Jogger	Courir	Boxer	Applaudir	Agiter	Moyenne	Écart type
Ke [44]	80,6%	36,1%	44,4%	69,4%	55,6%	91,7%	**62,9%**	19,6
Schuldt [77]	83,8%	60,4%	54,9%	97,9%	59,7%	73,6%	**71,7%**	15,2
Meng [60]	66%	62,5%	79,9%	88,8%	**93,1%**	91,7%	**80,3%**	12,2
Méthode 1	77,8%	**97,2%**	80,6%	80,6%	74,3%	77,8%	**81,4%**	**7,4**
Niebles [66]	79%	52%	88%	**100%**	77%	93%	**81,5%**	15,4
Kienzle [45]	95%	65%	71%	86%	91%	89%	**82,8%**	10,9
Méthode 2	**94,5%**	**97,2%**	**100%**	75%	37,1%	**97,2%**	**83,5%**	22,3

Tableau 2.17 – Les taux de bien classé et l'écart type selon chaque catégorie, pour différents travaux.

Cette comparaison confirme les performances de notre modèle. En effet, ce dernier classe correctement la plupart des catégories. Grâce à la combinaison CSST utilisée, notre modèle arrive à reconnaitre presque toutes les vidéos de la catégorie Jogger (taux de $97,2\%$). Contrairement aux autres travaux, qui ne classent pas plus de six sur dix. Malgré la complexité d'une telle catégorie, due à sa similarité avec les catégories Courir et Marcher, nous obtenons les meilleures performances. Avec le MBRL, nous obtenons une invariance dans les résultats de la classification des catégories d'actions humaines. Cette invariance est illustrée par le faible écart type obtenu $(7,4)$. Aucun des sept autres modèles présentés ne permet une telle stabilité. Pour la "méthode 2", la seule limitation réside dans la classification de la catégorie Applaudir.

Pour résumer, les caractéristiques que nous choisissons permettent de reconnaitre plus de huit vidéos sur dix, quelle que soit la catégorie d'actions humaines. Nous obtenons même une invariance des taux de bien classé, avec la "méthode 1", ce qui démontre la stabilité et la performance de nos résultats contrairement aux autres travaux. Notons que deux

nouveaux travaux ont été élaborés à la fin de la réalisation de ce mémoire [85, 50], les résultats fournis par les articles sont meilleurs de $3,2\%$ et $8,3\%$ que le nôtre. Le résultat de Wong et Cipolla n'est pas comparable au nôtre puisque les auteurs utilisent plus de données lors de l'apprentissage. Laptev *et al.* obtiennent pour la même base de données une moyenne de $91,8\%$ et ils utilisent la combinaison de la caractéristique du flou optique et de l'histogramme des orientations du gradient avec la classification par le SVM non linéaire.

2.5 Conclusion

Depuis plusieurs années, le problème de la reconnaissance d'actions humaines prend de plus en plus d'envergures. Plusieurs applications se sont basées sur des caractéristiques spatio-temporelles pour reconnaitre une action. Nous présentons une approche basée sur une nouvelle caractéristique CSST et un modèle de classification. La force de notre méthode est la combinaison de deux caractéristiques spatio-temporelles. La première étant les points d'intérêts situés dans des zones d'intérêts, appelés PIST (Points d'Intérêts Spatio-Temporels). À partir de cette zone, nous détectons le Contour Spatio-Temporel (CST). Ce dernier représente la deuxième caractéristique. Cette combinaison nous assure un grand nombre d'informations en un temps de calcul raisonnable. Après la réduction de données, notre classificateur basé sur le Modèle Bayésien de Régression Logistique (MBRL) est appliqué. Testés sur une des plus grandes bases de données d'actions humaines, nous avons présenté des résultats expérimentaux qui montrent que notre méthode est performante et robuste. D'ailleurs, une comparaison avec les autres travaux existants confirme ces performances, considérées comme les meilleures.

Plusieurs applications, telles que la vidéosurveillance intelligente ou l'indexation de vidéos, font appel à la reconnaissance d'actions humaines. Notre approche est utilisée dans

le prochain chapitre pour la réalisation d'un système d'interprétation des actions humaines pour la fabrication d'un dictionnaire de la vidéo. La MBRL que nous utilisons sépare linéairement les classes. Ce choix est motivé par la disponibilité du code. En perspective, l'utilisation de la MBRL à base de noyaux [72] pourrait améliorer la précision de la classification. Nous proposons aussi d'améliorer la détection des PIST. Cela est possible en éliminant les points n'appartenant pas au mouvement, et en ajoutant un algorithme d'appariement. Ce dernier permettrait d'avoir plus de précisions sur l'action. Il serait intéressant de généraliser notre modèle pour considérer des actions humaines non connues par le système. Cela pourrait se faire en ajoutant un module de mise à jour à notre approche. Une autre faiblesse de notre modèle est l'utilisation de l'ACP pour réduire nos données, il serait intéressant de le changer par un autre modèle statistique de réduction qui met en évidence la redondance des données.

CHAPITRE 3

Dictionnaire de la vidéo

3.1 Introduction

La production de vidéos connait une évolution spectaculaire. L'UNESCO estime dans son rapport ISU [5] que 1091 longs métrages 35mm ont été produits en Inde en 2006, 872 au Nigeria en 2005 et 485 aux États-Unis en 2006. Rien qu'au Québec, 100 longs métrages (de 60 minutes ou plus) ont été produit en 2007 [7]. De plus, les technologies d'acquisition, de visualisation et de diffusion de vidéos ont connu une grande avancée. Les caméras haute définition (HD), les écrans plats, les caméscopes numériques, le web, les cellulaires, les PDAs, la transmission satellitaire et internet haute vitesse sont des exemples qui témoignent de cette avancée. Ces technologies ont permis une prolifération de vidéos surtout sur le web. Par exemple, d'après le site de mediatedcultures [4], en mars 2008 le nombre de vidéos vues chaque jour sur Youtube est de 100 millions parmi un nombre total de 78,3 millions de vidéos existants sur le site.

La vidéo est utilisée principalement dans trois domaines d'application qui sont la diffusion d'information, la robotique et la sécurité. Considérons des exemples pour illustrer cette

utilisation dans ces trois domaines.

La vidéo est utilisée comme support pour la diffusion d'information (les films, les documentaires, etc.) dans le secteur de l'industrie cinématographique (DreamWorks SKG, Fireworks Pictures, etc.), de la télévision (CNN, Radio Canada, TVA, Euronews, etc.) et du web (Youtube, Dailymotion, etc.) où les usagers échangent et partagent d'innombrables vidéos. Dans ces secteurs, à part la fabrication et la diffusion des vidéos, d'énormes collections de vidéos doivent être stockées et archivées efficacement pour permettre un accès facile aux utilisateurs [1].

Dans la robotique, la vidéo est utilisée comme un outil pour effectuer certaines tâches. Par exemple, des robots utilisent la vidéo capturée et des techniques d'analyse d'images et de vidéos pour reconnaitre l'environnement qui les entoure afin d'éviter des obstacles, d'avancer, de reculer, de tourner, etc. [3]

Dans le domaine de la sécurité, la vidéo est utilisée comme un outil de prévention, de surveillance et aussi comme un outil d'aide à la prise de décision. Les systèmes de vidéosurveillance se basent sur la vidéo capturée par des caméras placées à des endroits stratégiques (routes, aéroports, stations de métro, etc.). Ces systèmes fonctionnent en temps réel pour identifier des événements suspects, des objets abandonnés ou même des actions non permises [2]. En plus des secteurs déjà cités, d'autres tels que l'éducation, l'industrie, la médecine, etc. intègrent de plus en plus la vidéo dans leur fonctionnement.

Malgré son considérable progrès, l'utilisation de la vidéo dans ces domaines d'application rencontre encore des limites, surtout lors de l'interprétation des événements et la génération automatique de la sémantique. Cette dernière se traduit par la reconnaissance du scénario, d'événements et des objets. Par exemple, dans le domaine de la sécurité, un système de vidéosurveillance tel que *IndigoVision* capte et identifie en temps réel des événements suspects ou des objets abandonnés [2]. Cependant, il n'est pas entièrement automatique puisqu'une personne doit l'assister lors de la prise de décision. Toutefois,

cette assistance humaine peut biaiser la décision à cause d'un manque de concentration, d'un oubli, etc. D'après Gouaillier et Fleurant, un surveillant ne peut pas suivre attentivement de 9 à 12 caméras plus de 15 minutes [31]. À ce phénomène s'ajoutent d'autres limites tel que le coût élevé de la main-d'œuvre.

Ces limites se retrouvent aussi dans le domaine de la diffusion d'information. Les sites web doivent permettre un accès rapide et efficace aux vidéos stockées et ceci par attribution de mots clés décrivant la sémantique de la vidéo. Cependant, l'homme intervient encore dans cette attribution ce qui peut induire des erreurs. Par exemple, dans la recherche de vidéos sur Youtube, il existe un manque de précision dans les résultats obtenus qui est dû, entre autres, à l'interprétation, l'humeur et le niveau intellectuel de l'utilisateur lors de l'attribution des mots clés à la vidéo au moment du téléversement. De même, pour la télévision, le stockage de vidéos peut faire face à des difficultés dues au coût élevé de l'analyse et de l'interprétation des événements de la vidéo.

Dans le domaine de la robotique, des limites relatives à l'invariance atmosphérique et à la prise de vue sont observées. Des robots comme *Spirit* et *Opportunity* effectuent certaines tâches (éviter un obstacle, avancer, etc.) en analysant et en interprétant l'information visuelle provenant de leurs neuf caméras ainsi que l'information apportée par d'autres capteurs. L'ensemble de ces informations leur permet d'interagir avec l'environnement sur la planète Mars [3]. Ces robots ne comportent pas d'outils d'analyse vidéo pour un environnement différent de celui de Mars pour lequel ils étaient prévus. Par exemple, le flou causé par la brume ou la pluie peut fausser l'interprétation de la vidéo et donc ces machines ne peuvent être utilisées que dans ce milieu particulier [56, 55].

Nous proposons dans ce chapitre un système d'interprétation de la vidéo pour la production de métadonnées en analysant et classifiant avec précision les événements de la vidéo sans intervention humaine. L'interprétation de ces événements est décrite à l'aide d'un rapport appelé « dictionnaire de la vidéo ». Comme le système est automatique et fournit

des informations précises, il peut réduire les erreurs relatives à l'intervention humaine. Ainsi, pour un système de vidéosurveillance, une concentration continue d'un surveillant ne sera plus exigée puisque ce dernier ne sera alerté que dans des cas suspects. De même, le stockage et l'archivage des vidéos deviennent indépendants des exercices manuels. Ce système, grâce à l'extraction automatique des métadonnées, permet également une uniformisation des mots clés qui se retrouvent dans les sites web contenant des vidéos. Le système peut aussi, vu sa rapidité, réaliser un gain de temps et par conséquent un gain du coût de la main-d'œuvre surtout celle spécialisée. Comme le dictionnaire fournit une description détaillée des événements de la vidéo, le système qui le génère peut avoir un autre avantage, celui d'être utilisé dans plus d'un domaine plutôt que d'être spécialisé dans un seul. Les métadonnées extraites dépendent généralement des données en entrée et donc peuvent changer selon les besoins d'un domaine en particulier. Dans cette perspective de formation des métadonnées, la méthode développée dans le deuxième chapitre pour la reconnaissance d'actions humaines est utilisée dans ce chapitre. Les hypothèses sur les vidéos et les actions sous-jacentes à notre approche sont : 1) Les vidéos sont non-compressés et à niveau de gris. 2) La caméra est fixe. 3) Les actions étudiés sont des actions humaines et dont les mouvements sont significatifs.

Ce chapitre est organisé comme suit. D'abord, nous décrivons dans la section 3.2 un système idéal pour la fabrication du dictionnaire. Puis, nous proposons dans la section 3.3 un état de l'art de toutes les procédures qui forment notre système. Dans la section 3.4, nous exposons notre Système d'Interprétation pour la Fabrication du Dictionnaire (SIFD) ainsi que les résultats obtenus et nous terminons avec une conclusion.

3.2 Caractéristiques d'un système idéal d'interprétation de vidéos

Notre travail a pour objectif la réalisation d'un système d'interprétation automatique pour la fabrication d'un dictionnaire de la vidéo. Ce dictionnaire est la description haut niveau des événements extraits de la vidéo. Selon notre analyse et notre discussion précédentes, nous définissons six critères qu'un système idéal d'interprétation vidéo doit satisfaire et qui sont : la généralisation, la précision, l'invariance, l'automatisation, la rapidité et le coût faible.

La généralisation du système se définit par son adaptabilité aux trois domaines d'utilisation de la vidéo déjà mentionnés dans l'introduction et à ses différents genres. Ainsi, lorsque ces deux exigences sont satisfaites, ce système peut interpréter de la même manière et avec la même précision aussi bien un film qu'une vidéo de surveillance, une vidéo compressée qu'une vidéo non compressée, une vidéo en couleur qu'une vidéo à niveaux de gris.

La précision se définit comme une mesure du degré avec lequel les résultats produits par le système sont conformes à la vérité terrain. Le système est précis lorsque, appliqué à une vidéo, il parvient à détecter les plans, les objets et le mouvement avec un maximum de précision. Si un événement est non détecté ou rajouté ou mal interprété, cela peut changer le sens du scénario de la vidéo ainsi que la sémantique qu'elle contient. La précision exige pour cela une augmentation des vrais positifs, des biens suivis et des biens reconnus et exige une diminution des faux négatifs, des faux positifs, des non suivis, des mals suivis et des non reconnus. Un vrai positif est un objet existant dans la vidéo et bien détecté par le système ou un plan bien déterminé. Un faux négatif est un objet qui existe dans la vidéo, mais ignoré par le système ou un plan ignoré par le système. Les faux positifs sont les objets trouvés par le système et non existants dans la vidéo ou les plans détectés

par le système et n'existant pas dans la vidéo. Un objet est bien suivi est un objet dont la trajectoire trouvée par le système est celle existante dans la vidéo. Un objet non suivi est un objet dont la trajectoire n'a pas pu être déterminée. Un objet mal suivi est un objet dont la trajectoire n'a pu être que partiellement déterminée. Un objet bien reconnu est un objet dont l'action effectuée est bien identifiée par le système. Enfin, un objet mal reconnu est un objet dont l'action déterminée par le système ne coïncide pas avec celle de la vidéo.

Un système invariant se définit généralement par une utilisation stable et précise, indépendamment des conditions de changements de point de vue ou atmosphériques. Par exemple, la reconnaissance des actions d'un personnage dans un film doit être précise dans les différentes conditions atmosphériques pour que le scénario de l'acteur reste cohérent. Elle est aussi importante dans la robotique afin que des robots comme *Spirit* et *Opportunity* distinguent les objets lors de leur mouvement même avec des changements de luminosité sur Mars ou même sur Terre.

Aussi, un système idéal se définit comme étant automatique en réduisant, dans les étapes du processus de fabrication du dictionnaire, l'intervention humaine. L'utilisateur n'intervient presque pas ni dans la reconnaissance d'événements ni dans celle des objets. Cet objectif permet de diminuer surtout le coût de l'exploitation de la vidéo. Par exemple, Youtube peut indexer automatiquement les vidéos.

Idéalement, un système de production de dictionnaire doit fonctionner en temps réel. Il est possible de réaliser un tel système en utilisant une grappe de calcul. Le défi consiste à implanter un tel système en utilisant le minimum de noeuds pour réduire les coûts d'exploitation. Un compromis entre le coût d'exploitation et le temps d'exécution doit être atteint.

Pour satisfaire ces exigences, nous pensons qu'une architecture en trois étapes peut être

développée. Premièrement, une identification du genre de la vidéo est requise, à savoir si elle est compressée ou non, quel est le mode de compression (MPEG, MPEG2, WMV, etc.), quel espace couleur est utilisé (RGB, HSI, niveaux de gris,...), etc. Par cette distinction, les procédures d'extraction des événements de la vidéo sont mieux adaptées. Par exemple, pour une vidéo compressée MPEG, seulement le *bitstream* peut être utilisé dans le reste des procédures du système. Deuxièmement, les objets, les trajectoires, les plans d'une vidéo, etc. sont séparés et distingués afin de garder les données pertinentes et alléger les procédures de reconnaissance d'événements. Par exemple, les vecteurs de mouvement de MPEG peuvent être utilisés. Troisièmement, les résultats obtenus par ces procédures sont fusionnés pour mettre en relation les événements détectés et donc former le dictionnaire. D'autres sources de données de la vidéo peuvent être associées aux images, comme la voix et le texte (sous-titrage).

3.3 État de l'art

Pour élaborer un dictionnaire de la vidéo, il est nécessaire de représenter cette vidéo. Il existe trois niveaux de représentation d'une vidéo. La représentation *bas-niveau* qui décrit les caractéristiques d'un contenu vidéo par des outils de base tels que la couleur, la texture, les formes et les mouvements. Puis, il existe la représentation *structurelle* qui décrit une organisation structurelle de la vidéo en images, plans, scènes et séquences comme l'illustre la figure 3.1. Un plan est défini comme une séquence d'images prises par une seule caméra stable en continu. Les scènes sont définies comme des suites de plans contigus qui sont sémantiquement reliés. L'action est effectuée dans un même environnement et dans un temps continu. Une séquence est alors un ensemble de scènes appartenant au même élan de narration et d'émotion cinématographique. Cette représentation est généralement utilisée dans la production cinématographique. La dernière représentation est celle de

haut-niveau. Elle fournit une description sémantique du contenu de la vidéo dans le but de modéliser « l'histoire » véhiculée dans la vidéo [41].

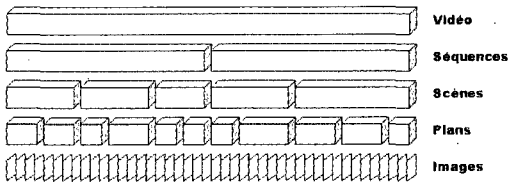


Figure 3.1 – Structure cinématographique d'une vidéo.

La représentation *bas-niveau* cherche à décortiquer les éléments importants dans une vidéo, essentiellement les objets en mouvement et leur trajectoire. Une des techniques établies pour trouver les objets est la segmentation, tandis que le suivi d'objets est la méthode nécessaire pour avoir la trajectoire des objets segmentés. Nous effectuons, dans ce qui suit, une revue des techniques de segmentation et de suivi d'objets dans la vidéo.

La segmentation a pour but l'extraction de l'information pertinente de la vidéo. Dans les images fixes, la segmentation a pour but de distinguer les différentes régions. Par contre, pour la vidéo, la segmentation est l'extraction des objets en mouvement. Il existe trois principales approches de segmentation. La première se base sur les différences d'images en considérant la caméra fixe. Cette segmentation est surtout utilisée dans les modèles dédiés à la vidéosurveillance. Elle comporte deux méthodes, une basée sur la différence avec l'image de référence et l'autre sur la différence entre images successives. Le principal problème de ce type de méthode est l'acquisition de l'image de référence et sa mise à jour à cause des effets surtout de changements d'illumination ainsi que le problème d'instabilité de la caméra. Lorsque l'image de référence est difficile à acquérir, elle est estimée [14] ou selon les travaux de Hsu et Tsan [39] l'objet mobile de la vidéo est

enlevé. Dans l'autre méthode, la différence d'images successives détecte les objets et surtout ceux qui effectuent un mouvement rapide, mais elle est très sensible au bruit [39]. La deuxième approche de segmentation, appelée construction des mosaïques, consiste dans le cas d'un mouvement latéral ou vertical de la caméra à estimer le fond et par conséquent discriminer les objets présents dans le premier plan. Le mouvement global entre deux images successives est estimé et une seule image panoramique est composée par ces deux images [76, 82, 40]. La dernière approche est la segmentation de mouvement où les images sont découpées en régions ayant un mouvement homogène. Pour cette fin, les paramètres du mouvement sont estimés en utilisant la fusion [17], ou le test bayésien [18], ou l'agrégation autour de centres mobiles [8].

Le suivi d'objets en mouvements est une technique pour la représentation *bas-niveau* de la vidéo. Il existe un grand nombre d'approches pour effectuer le suivi. Yilmaz *et al.* [88] classifient ces approches en trois catégories. Le suivi des points est la première catégorie. Il consiste à faire correspondre les points entre eux d'une image à l'autre. Il est nécessaire pour que la méthode fonctionne de détecter les bons points d'intérêts d'où la présence de certaines contraintes pour mieux les cibler. Sethi et Jain [79] utilisent une approche *gloutonne* avec des contraintes de proximité et de rigidité. Veenman *et al.* [83] ajoutent une nouvelle contrainte liée au mouvement pour pouvoir gérer les problèmes d'occultations rencontrés dans le travail précédent. Schmid et Mohr [78] ajoutent une contrainte géométrique dont le principe est de retrouver les mêmes points voisins de l'image d'origine dans l'image cible. Une contrainte de triangulation de Delaunay est ajoutée dans les travaux de Remi et Bernard [74]. La deuxième catégorie est celle de suivi de silhouettes. La méthode des contours actifs est un exemple. Elle consiste à faire évoluer le contour initial vers l'objet d'intérêt [80, 24]. En minimisant la distance Kulback-Leibler (KL) entre les statistiques locale et globale de l'objet par rapport au fond, le suivi est effectué en présence de contraintes de proximité et de rigidité. Allili et Ziou [11] formulent la

région d'intérêt, pour gérer les problèmes d'occultations évoqués dans le travail de Sethi et Jain [79]. Une autre catégorie est le suivi de noyau. Elle consiste dans le suivi d'une forme géométrique basique telle qu'un rectangle ou une ellipse. Les chercheurs utilisent la couleur et le gradient pour effectuer une recherche exhaustive du même objet dans l'image suivante.

La représentation *structurelle* de la vidéo décrit l'organisation de la vidéo. Plusieurs travaux ont été réalisés sur cette représentation surtout celle de la détection de plans. Cependant, les techniques de transition, qui sont de plus en plus nombreuses et variées, restent encore à étudier. Il est possible de diviser les changements de plan en deux genres : les changements progressifs définis par une continuité visuelle lors du passage d'un plan à l'autre et les changements brusques qui consistent à passer d'un plan à un autre sans transition. Il est facile de détecter ce dernier genre de transition. Par contre, les changements progressifs sont plus difficiles à détecter. Il existe principalement cinq différentes caractéristiques pour la détection de plan. Premièrement, les méthodes orientées pixels que les différences d'intensité, par exemple, illustrent. Une des plus simples techniques dans ce domaine est celle de la différence d'intensité moyenne développée par Nagasaka et Tanaka [63]. Aigrain et Joly [10] proposent aussi une méthode basée sur une étude statistique des différences pixel à pixel. Deuxièmement, les méthodes orientées histogrammes qui présentent des méthodes de différences et de comparaison d'histogrammes [65]. Troisièmement, les méthodes orientées contours ont des avantages surtout pour leurs invariances aux changements d'illumination. Les travaux de Nam et Tewfik, de Zabih *et al.* et de 200 [64, 89, 1] illustrent ce genre de méthodes. Quatrièmement, les méthodes orientées transformées de coefficients telles que la transformée de Fourier discrète, DCT et ondelette sont aussi utilisées. Cinquièmement, les méthodes orientées mouvement sont utilisées. Elles sont généralement combinées à d'autres caractéristiques pour la détection de plan et elles sont généralement inefficaces en cas d'absence de mouvement. Zhang

et al. utilisent une technique de seuillage pour identifier les endroits où les changements (mouvement) sont de faible amplitude. D'autres méthodes se basant sur les sciences cognitives et le traitement du signal sont de plus en plus développées [92]. Boccignone *et al.* proposent une technique de découpage en plans fondée sur la focalisation de l'attention issue du système visuel humain [16].

La dernière représentation est celle de *haut-niveau.* Lors de cette représentation, la détection d'événements est effectuée. Elle consiste à rechercher des relations entre les mouvements et donc à créer la sémantique de la vidéo. Il existe trois catégories en général pour la détection d'événements. La première est celle des approches basées sur des modèles d'événements prédéfinis par des formes, des règles et des contraintes. L'approche de Koller *et al.* [47] utilise les verbes de mouvement pour décrire les événements effectués par des voitures, des autobus, etc. par exemple « arrêt » ou « avancer », etc. Babaguchi et Jain [12] se basent sur la collaboration intermodale qui utilise le domaine de connaissance basé sur les mots clés d'événements. Une autre méthode utilise la segmentation du mouvement spatio-temporel [75]. Ben-Arie *et al.* [13] utilisent la vitesse des vecteurs des parties du corps et l'indexation de pose. Il existe aussi une méthode d'image fondée sur l'ontologie de reconnaissance développée par Maillot *et al.* [58]. D'autres encore utilisent la taxonomie et l'ontologie [33]. La deuxième catégorie est celle des approches qui apprennent automatiquement les événements au lieu de les spécifier manuellement, et cela, en utilisant des données d'apprentissage [19, 30, 38, 42]. La troisième catégorie se base sur les méthodes de classification d'événements utilisant les techniques de (*clustering*) [73, 90, 93].

Les recherches effectuées s'intéressent à combiner plusieurs représentations de la vidéo. Ces représentations ont pour but de décrire automatiquement les événements de la vidéo. Par exemple, dans le travail de Hakeem et Shah, un roman de la vidéo est fabriqué où ils développent un modèle de lecture, de détection et de représentation d'événements dans la

vidéo. C'est dans cette optique que notre dictionnaire se situe en intégrant les différentes représentations de la vidéo dans un seul système.

3.4 Le dictionnaire

3.4.1 Objectif

Pour élaborer ce dictionnaire, notre système doit répondre au mieux aux exigences d'un système idéal, déjà présenté dans la section 3.1. Cependant, un tel système est difficile à réaliser surtout avec les contraintes de la multitude de leurs formats vidéo, de leur qualité et de leur contenu. Nous choisissons dans notre travail d'émettre certaines hypothèses dans l'objectif de garder une vision générale sur sa précision et son automatisme. L'autre objectif de ce travail est de tester l'utilisation de notre dictionnaire dans deux différents types de vidéos.

3.4.2 Caractéristiques du Système d'Interprétation pour la Fabrication du Dictionnaire (SIFD)

La multitude d'informations possible à extraire et les différents genres d'objets qui peuvent se retrouver dans une vidéo rendent un système de détection d'événements généralement difficile à réaliser, surtout avec la contrainte de précision. Afin de faciliter la réalisation d'un système de haute précision, nous appliquons des hypothèses sur les vidéos à étudier. Le SIFD est appliqué sur des vidéos non compressées captées par des caméras fixes. Notre système reconnait seulement les actions humaines et ne considère pas les relations qui peuvent exister entre elles. Le SIFD produit un dictionnaire en sortie détaillant tout ce qui se passe dans une vidéo : les objets, les mouvements humains

effectués, le début et la fin d'un mouvement et le début de chaque plan.

La figure 3.2 présente les quatre principales procédures pour l'élaboration du dictionnaire. Au départ, une procédure de détection de coupures de plan est effectuée. À partir de cette dernière, une première information qui est l'image de coupure ainsi que sa position est acheminée au dictionnaire. Les plans résultants avec un certain nombre d'actions à détecter constituent l'entrée pour la segmentation. Ensuite, une segmentation par le contour spatio-temporel est effectuée pour chaque séquence d'images d'un même plan. À la fin de la segmentation, une information sur la forme des objets et la trajectoire du mouvement est ajoutée au dictionnaire. Cette information est fournie en entrée pour la troisième procédure, celle de la reconnaissance d'actions humaines. Un dictionnaire est ainsi formé par toutes les coupures, les objets dans chaque plan et le mouvement que ces objets effectuent. Nous détaillons en ce qui suit chacune des procédures pour former le dictionnaire. La reconnaissance d'actions humaines fera l'objet du chapitre 2 de notre mémoire et le reste des procédures du dictionnaire seront au chapitre 3.

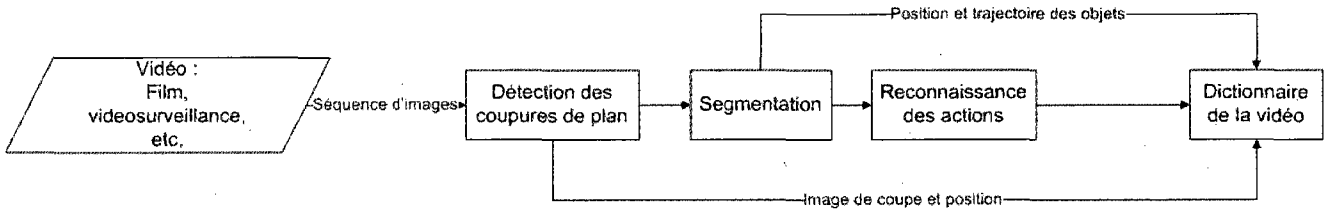


Figure 3.2 – SIFD système d'interprétation pour la fabrication du dictionnaire

3.4.3 Détection des plans

Un plan est l'unité de base dans la construction d'une vidéo. C'est une suite ininterrompue d'images provenant d'un seul enregistrement d'une caméra [51]. Il existe différents

types de coupures dont les plus utilisées sont la coupe franche, le fondu en ouverture, le fondu en fermeture, le fondu enchainé et le volet. Une coupe franche consiste à passer instantanément d'un plan à l'autre et produit des changements visuels brusques. Les fondus en ouverture et en fermeture sont des changements contrôlés de l'amplitude du signal image. Un fondu en ouverture fait progressivement apparaître un plan à partir d'une image noire, tandis qu'un fondu en fermeture fait disparaître un plan vers une image noire. La quatrième coupe, le fondu enchaîné (*dissolve*) est une combinaison d'un fondu en ouverture et d'un fondu en fermeture. Ce genre de coupure est utilisé pour signifier un saut dans le temps. Enfin, le volet (*wipe*) est une transition graduelle à l'aide d'un trucage vidéo par lequel une image semble pousser une autre en dehors de l'écran. Dans notre travail, nous utilisons la méthode développée au sein du laboratoire MOIVRE par Lawrence *et al.* [51]. Ce choix est motivé par la disponibilité du code source. Cette méthode, appelée détection de frontières de plans, se base sur la sommation de la dérivée temporelle et de la suppression des pixels où le gradient n'est pas dominé par sa composante temporelle. Lawrence *et al.* définissent une mesure $D(t)$ pour la segmentation de vidéo 3D. Pour une vidéo $V(x, y, t)$, ils calculent le gradient spatio-temporel à l'aide des trois dérivées partielles $V_x(x, y, t)$, $V_y(x, y, t)$ et $V_t(x, y, t)$. La mesure $D(t)$ est calculée alors par l'équation

$$D(t) = \sum_{x=1}^{M} \sum_{y=1}^{N} \begin{cases} |V_t(x, y, t)| & \text{si } \tan\theta < \tau_{mvt} \\ 0 & \text{sinon} \end{cases}$$

où M et N sont respectivement la longueur et la largeur des images de la vidéo, τ_{mvt} est un seuil défini à l'entrée et θ est l'angle entre le vecteur du gradient (V_x, V_y, V_t) et le vecteur temporel $(0, 0, V_t)$ tel que

$$\tan\theta = \frac{\sqrt{V_x^2 + V_y^2}}{|V_t|} \tag{3.1}$$

Les dérivées partielles sont développées dans la section 2.1.2. Les coupures produisent des maxima dans la mesure $D(t)$. Pour localiser alors les frontières, il suffit d'identifier

les maxima locaux et supprimer ceux qui peuvent être dus à des phénomènes comme le bruit, le mouvement et les variations dans l'éclairage. Les frontières sont localisées telles que

$$t \text{ est un max local } \Leftrightarrow D(t) > D(s) \quad \forall s \in [t-\omega, t+\omega] \tag{3.2}$$

où $2\omega + 1$ est la taille du voisinage de t considéré pour la localisation de maxima. Ces derniers sont analysés pour éliminer ceux non désirés. Les maxima de faible amplitude sont éliminés par un seuil τ_1. Puis, parmi le reste des maxima, ceux avec une faible amplitude relative à leur voisinage sont aussi éliminés. L'amplitude relative AR est définie comme

$$AR(t_0) = D(t_0) - \frac{(m_a + m_b)}{2} \tag{3.3}$$

où m_a et m_b sont respectivement les valeurs minimales de la fonction $D(t)$ avant et après le maximum $D(t_0)$ dans une fenêtre de taille fixe $4\omega + 1$ centrée sur t.

$$m_a = Min(D(s)) \quad \forall s \in [t_0 - 2\omega, t_0 - 1] \tag{3.4}$$

$$m_b = Min(D(s)) \quad \forall s \in [t_0 + 1, t_0 + 2\omega] \tag{3.5}$$

La frontière t_0 de la vidéo est identifiée comme frontière d'un plan si

$$D(t_0) \text{ est un max local de D(t)} \tag{3.6}$$

$$D(t_0) > \tau_1 * Max(D(t)) \; \forall t \tag{3.7}$$

et

$$D(t_0) > \tau_2 * AR(t_0) \tag{3.8}$$

3.4.4 Extraction des zones d'intérêts des vidéos

Après l'extraction des images de coupures, nous obtenons alors un certain nombre de séquences d'images homogènes et contenant un certain nombre d'objets. Nous détectons

ces objets de la séquence. Pour ce faire, nous procédons par la détection du contour spatio-temporel. Tel qu'il a été présenté dans la section 2.1.2, le contour spatio-temporel contient l'information spatiale des objets en mouvement. Nous considérons comme objet spatio-temporel, une zone de l'image contenant un contour spatio-temporel avec les pixels contigus. Ainsi, ces contours vont être utilisés comme caractéristique CST lors de la reconnaissance de l'action humaine pour générer les métadonnées dans notre dictionnaire.

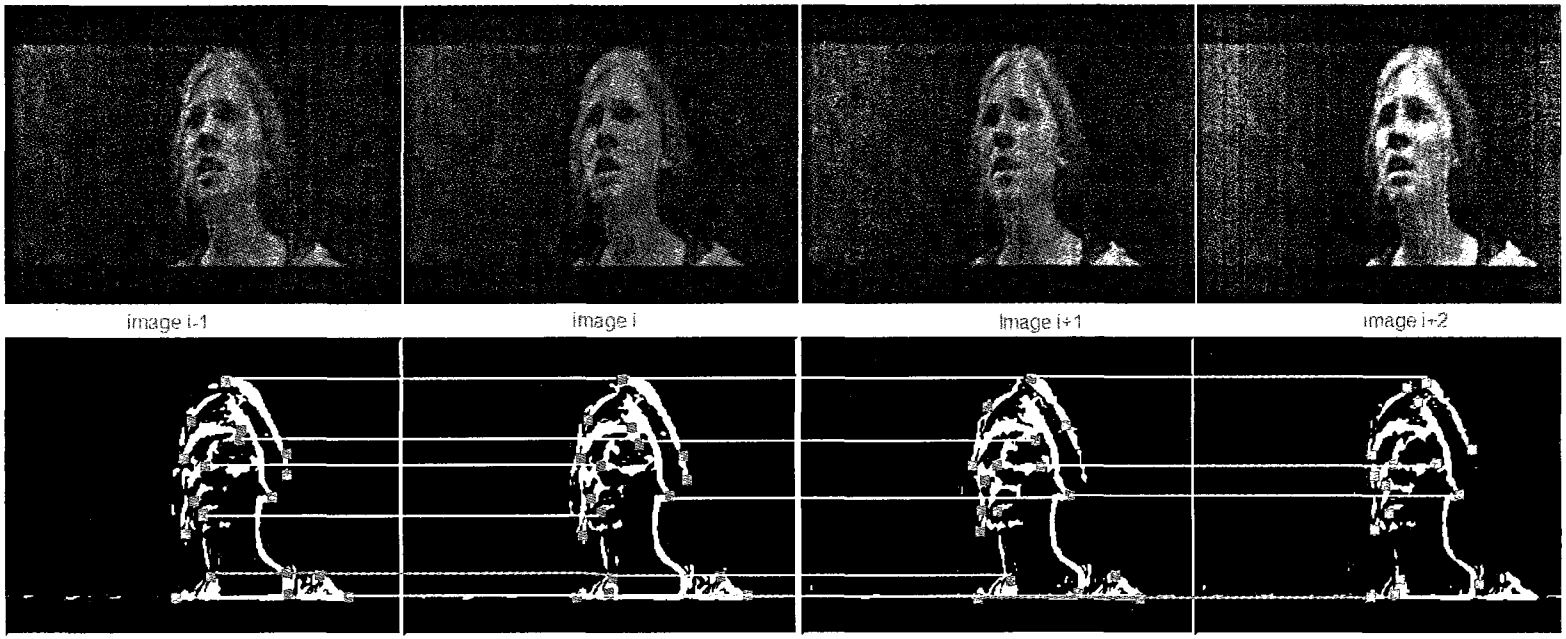


Figure 3.3 – Suivi d'une personne qui bouge la tête

3.4.5 Suivi des zones d'intérêts

Pour chaque plan trouvé dans la vidéo, un ensemble de zones sont extraites par le contour spatio-temporel. Nous suivons après le mouvement de ces zones et nous déterminons la trajectoire qu'elles effectuent. Le suivi fait correspondre les zones extraites et sépare entre plusieurs objets en mouvement. Pour suivre une zone 1 trouvée dans l'image $i-1$, nous détectons les points d'intérêts SIFT S_{i-1}^1 qu'elle contient. Nous cherchons pour chaque zone détectée j de l'image i du gradient spatio-temporel les points SIFT S_i^j qu'elle contient. Parmi les points S_{i-1}^1 et S_i^j, nous cherchons ceux qui correspondent en coordonnées. La zone j de l'image i du gradient ayant le plus de points correspondants est considérée comme la zone 1 en mouvement. Nous effectuons la même procédure pour le reste des zones de l'image i du gradient. Une zone est considérée disparue si aucun point n'est retrouvé dans l'image suivante. Une nouvelle zone est crée si une zone est détectée et ne correspond à aucune zone antérieure. Nous illustrons ce suivi d'objets à la figure 3.3.

Nous poursuivons les zones image par image jusqu'à la fin du plan. Lors de la fabrication de notre dictionnaire, nous décrivons la trajectoire de l'objet pendant tout le plan. Il peut arriver qu'une zone, lors de la détection d'un contour, soit découpée en zones non contiguës. Par notre suivi, nous pouvons reconstruire la zone dans les images suivantes. Dans le cas où une zone M de l'image i se divise en deux morceaux M_1 et M_2 lors de l'extraction des zones d'intérêts de l'image $i+1$ et se reconstitue dans l'image $i+2$ en une seule zone M, notre système SIFD considère que M_1 et M_2 constituent un seul objet. Nous illustrons ce cas à la figure 3.4 où même si "une personne" se découpe en deux morceaux dans i, elle est retrouvée dans $i+1$. Notre système rassemble les deux morceaux de i lors de la correspondance de points de chaque zone dans l'image $i+1$, ces

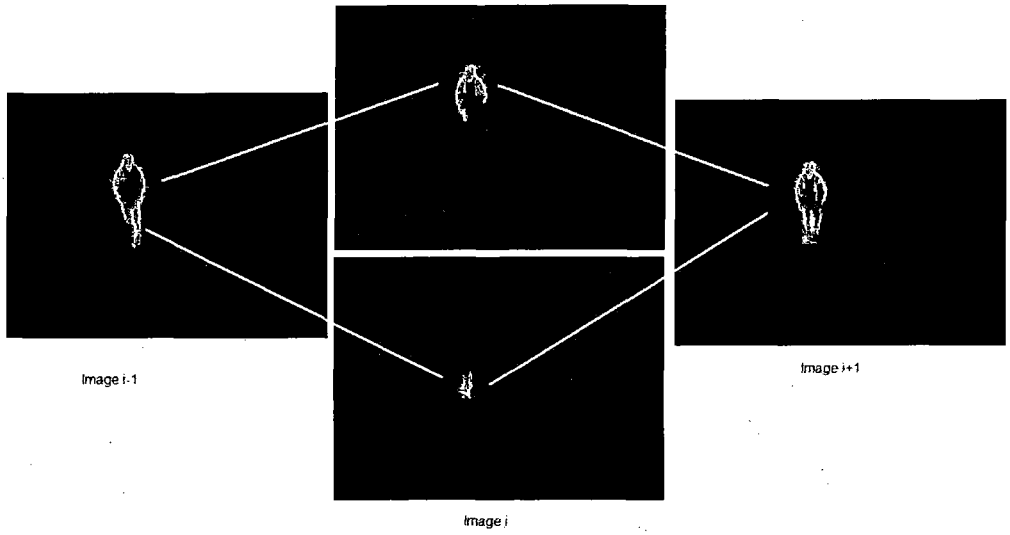


Figure 3.4 – Reconstruction d'objet par le suivi

deux dernières appartenant au même objet.

3.4.6 Reconnaissance d'actions humaines

Après l'extraction des zones et le suivi de leurs mouvements, nous reconnaissons le mouvement effectué par chaque objet. Dans notre travail, nous ne considérons que les actions humaines qui sont importantes que ça soit pour la vidéosurveillance ou pour les films et même pour les journaux télévisés. Nous appliquons le modèle de reconnaissance d'actions humaines développé dans le chapitre 2. Rappelons que la zone d'intérêt représentée par la caractéristique CSST qui englobe le contour spatio-temporel ainsi que les points d'intérêts déjà trouvés lors de l'extraction des zones d'intérêts et de suivi des objets. Les données extraites dans cette zone seront utilisées pour la reconnaissance de la classe d'actions.

Pour la procédure d'apprentissage, le système apprend des actions définies par l'utili-

sateur et contenues dans différentes vidéos. Ces vidéos contiennent chacune d'elles une seule action humaine. Lors de l'apprentissage, chaque ensemble de vidéos contenant la même action est regroupé. Comme présenté dans le chapitre 2, pour chaque vidéo d'apprentissage, le CSST est extrait seulement sur les zones d'intérêts. Nous apprenons nos actions par l'algorithme MBRL présenté dans la section 2.3. L'apprentissage pour la reconnaissance d'actions humaines est effectué sur seulement les actions provenant de notre collection et donc seulement les actions « marcher », « courir », « ouvrir la porte », « s'asseoir » et « se tenir debout » peuvent être reconnues pour le test. Pour chaque zone extraite et suivie dans un plan, nous lui fabriquons le vecteur de caractéristique CSST (section 2.1). Tel que présenté dans la section 2.3.3, nous définissons la classe de l'objet en mouvement de la zone d'intérêt par ce vecteur. L'avantage d'un système basé sur l'apprentissage est qu'il pourrait être utilisé dans plusieurs applications. Par exemple, si le système apprend les actions « courir », « frapper », « marcher », « sauter », le SIFD peut détecter une bagarre dans une station de métro par l'action « frapper ».

3.4.7 Expérimentations

Afin de montrer les caractéristiques du SIFD, nous présentons, au début, des mesures de performance pour chacune de ses procédures pour mettre en évidence sa précision et son automatisation. Ensuite, nous analysons à la fois une vidéo provenant d'un film et une autre provenant d'un système de vidéosurveillance. Avant de présenter les résultats obtenus, nous décrivons, en premier lieu, la définition de la vérité terrain et les abréviations utilisées au sein de notre dictionnaire. En deuxième lieu, nous comparons les résultats de la détection de plan, de l'extraction des zones d'intérêts, du suivi et de la reconnaissance d'actions humaines avec la vérité terrain. Les résultats de la reconnaissance sont aussi comparés avec ceux obtenus dans le troisième chapitre du mémoire de Chahid[21]. En troisième lieu, nous montrons les résultats obtenus du SIFD d'un épisode de la sé-

rie *Friends* (figure 3.10) et nous comparons ceux d'une vidéo provenant de la collection PETS au nom de *OneStopNoEnter1cor* avec sa vérité terrain.

Définition et abréviations

Les abréviations suivantes sont utilisées pour décrire les résultats obtenus : NM= nom du fichier de la vidéo, NB= Nombre d'images dans la vidéo, IW= largeur des images de la vidéo, IH= longueur des images de la vidéo.

Dans un plan, les abréviations suivantes sont utilisées : Pl= numéro du plan, IR = image de référence d'un plan, ID= numéro de l'image de départ d'un plan, IF= numéro de l'image finale d'un plan.

Pour les objets détectés dans une séquence d'images d'un même plan, les abréviations suivantes sont utilisées : O= numéro de l'objet détecté, IO= image de l'objet, OD= numéro de la première image où l'objet est détecté, OF= numéro de l'image de disparition de l'objet, Mv= l'action humaine effectuée.

Tel que définit dans la section 3.2, pour illustrer la précision de notre système, nous représentons le nombre d'objets **NB** dans la vérité terrain **OR**, le nombre **NB** et le taux **T** de vrais positifs (plans ou objets) estimé **V.P** par le système SIFD, le nombre **NB** et le taux **T** de vrais négatif **V.N** (plans ou objets), le nombre **NB** et le taux **T** de faux positifs **F.P** (des objets détectés non existants ou des plans non existants), le nombre **NB** d'objets à suivre **O.S**, le nombre **NB** et le taux **T** de bien suivi **B.S**, de non suivis **N.S** et de mal suivi **M.S**, les actions reconnues dans la vérité terrain **A**, le nombre **NB** et le taux **T** de biens reconnu **B.R** et enfin le nombre **NB** et le taux **T** de mals reconnus **M.R** des actions mal définies.

La vérité terrain

La vérité terrain est la reproduction manuelle de la détection de changement de plan, de la détection d'objets, du suivi d'objets et de la reconnaissance d'actions que nous effectuons sur une vidéo. En premier lieu, nous cherchons où se trouvent des changements de plan dans une vidéo. Pour les vidéos structurées, nous considérons les mêmes genres de plan que dans la littérature. Nous cherchons alors un changement brusque ou un changement progressif dans la vidéo et nous marquons l'image du début du changement. Pour les vidéos non structurées, nous considérons qu'un changement de plan est un changement significatif dans lequel un objet de taille supérieure à 50 pixels apparait dans plus de 10 images. Par exemple, dans une vidéo de surveillance, une personne qui entre dans le champ de la caméra de surveillance provoque un changement de plan. Nous marquons dans ce cas l'image de l'apparition de l'objet en tant qu'image de référence pour le nouveau plan.

En deuxième lieu, nous détectons les objets en mouvement. Nous visualisons image par image un plan et nous cherchons les objets en mouvement pour les encadrés par un rectangle dessiné par la souris de façon qu'il englobe l'objet en mouvement. Nous effectuons cette démarche jusqu'à la fin du plan. Dans la vérité terrain, nous considérons une zone de changement toute zone continue dans l'espace et en mouvement. Par exemple, nous considérons dans la figure 3.5 la zone de changement en rectangle rouge. Elle entoure une personne qui ouvre une porte et une partie de la porte qu'elle ouvre.

En troisième lieu nous suivons nos zones de changements image par image. Dès qu'une zone de changement est détectée dans l'image i, nous cherchons dans l'image $i + 1$ la zone qui lui correspond. Une zone qui se retrouve d'une image à l'autre est une zone dont la forme visuelle et la position ne varient pas beaucoup. Dans la figure 3.5, nous considérons que les trois zones de changements sont la même zone en mouvement. Nous considérons une zone est bien suivi **B.S** lorsqu'elle est détectée et suivie tout au long de

sa trajectoire dans un même plan. Une zone mal suivi **M.S** est celle où seulement une partie de sa trajectoire est conforme à celle de la vérité terrain.

En quatrième lieu, nous effectuons la reconnaissance de l'action trouvée après le suivi d'une zone. Nous attribuons le verbe d'action d'une zone selon les verbes donnés en apprentissage pour notre SIFD (marcher, ouvrir une porte, courir, etc.). En cas d'absence parmi nos verbes, nous attribuons un verbe à l'action n'appartenant pas au groupe déjà connu par exemple « sauter », « tomber », etc.

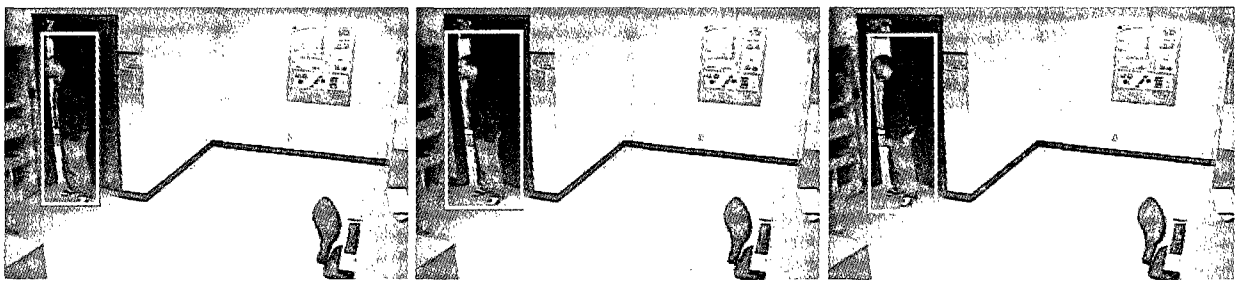


Figure 3.5 – Zone de changement extraite par la vérité terrain

La collection

Nous expérimentons notre algorithme sur trois collections de vidéos différentes, comportant des données du secteur de la surveillance, du cinéma et de la prévention. Premièrement, les vidéos de la collection PETS utilisées par un grand nombre de chercheurs dans les méthodes de segmentation et de classification de vidéo [6]. Ces vidéos MPEG comportent des actions qui se déroulent principalement dans un centre d'achat ou une station de métro. Dans notre travail, nous choisissons 15 vidéos de cette collection au hasard avec chacune comportant un échantillon de 16 secondes avec 25 images par seconde. Les images de ces vidéos sont de taille 320×240 pixels. Ensuite, une vidéo de la série *Friends* est extraite. Elle comporte plusieurs plans et des objets différents. Nous avons

choisi une partie de l'épisode de *Friends* de 600 secondes avec 25 images par seconde. Chaque image est de l'ordre de 160×120 pixels. La troisième collection est une collection de vidéos d'actions humaines que nous avons captées. Nous avons filmé 34 vidéos pour l'action « marcher », 22 pour celle de « courir », 29 pour l'action « ouvrir une porte », 13 « se tenir debout » et 14 vidéos pour l'action « s'asseoir » sur une chaise pour un total de 112 vidéos. Chaque vidéo comporte 200 images de 320×240 pixels. La figure 3.6 illustre notre collection. Nous effectuons l'apprentissage de notre système de reconnaissance d'actions humaines sur 66 vidéos parmi notre collection et nous les testons sur l'ensemble des vidéos de notre collection. L'action "se tenir debout" est une action où la personne est debout tout en effectuant un mouvement du corps (se retourner, se balancer de droite à gauche, etc.) sans effectué un déplacement dans la scène. Dans le but de montrer la précision du SIFD, nous disposons d'une vérité terrain obtenu manuellement.

Détection de plans

Nous comparons les résultats de la détection de changement de plan obtenus par SIFD avec ceux de la vérité terrain. Les résultats obtenus pour 15 vidéos de la collection PETS ainsi que ceux obtenus pour les 30 vidéos de notre collection sont présentés dans le tableau 3.1. Nous retrouvons, dans ce dernier, le nombre de transitions de plans à détecter pour chaque groupe de vidéo selon la vérité terrain ainsi que celui des vrais positifs, des faux positifs et des vrais négatifs estimés par SIFD. Ce tableau donne une idée sur les performances de la détection de plan dans le cas des vidéos non structurées (vidéo surveillance). Nous remarquons que, pour les vidéos de PETS et celles de notre collection, le nombre de **F.P** est élevé, ce qui peut être expliqué essentiellement par le choix des paramètres en entrée pour l'algorithme. Nous remarquons aussi un nombre élevé de **V.P** avec un taux de changements de plan bien détecté d'environ 98.5 % pour

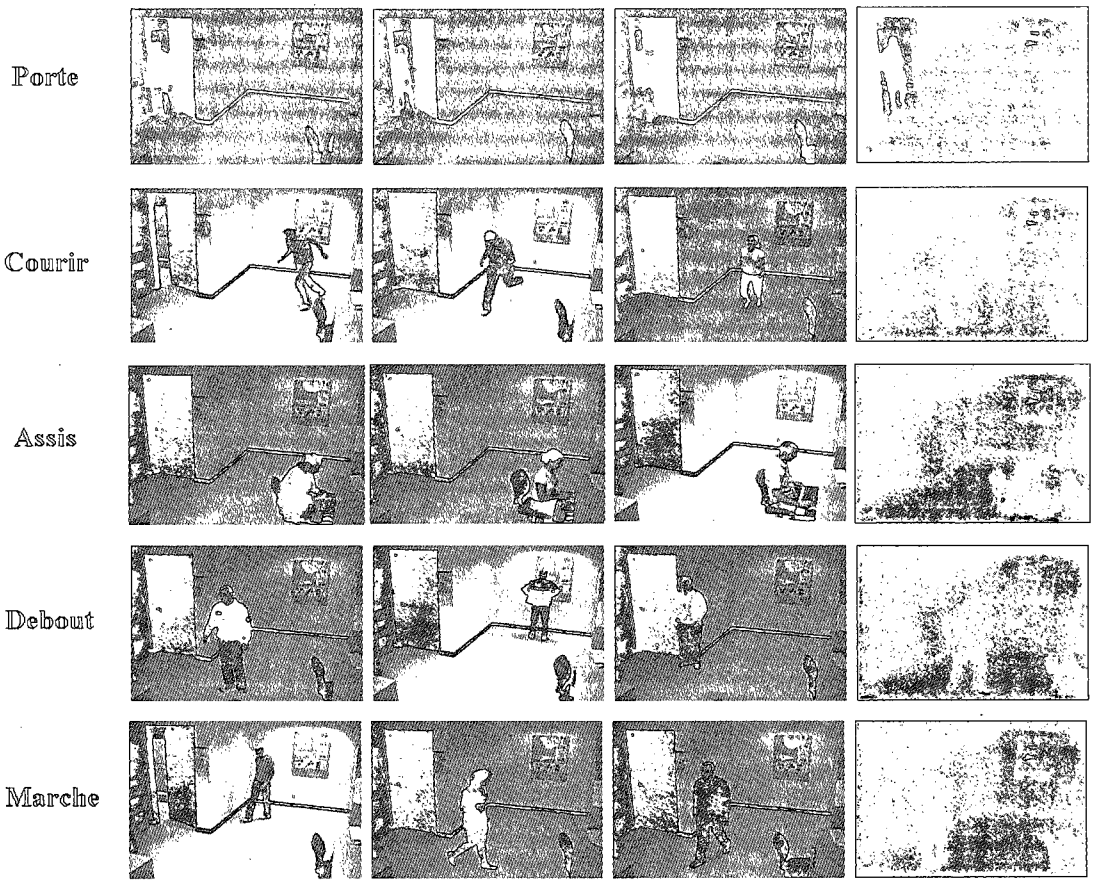


Figure 3.6 – Collection des actions humaines (marcher, courir, se tenir debout, ouvrir une porte et s'asseoir sur une chaise)

Titre Vidéo	OR	V.P	V.N	F.P
15 vidéos PETS	38	36	2	6
112 vidéos propre	85	85	0	10

Tableau 3.1 – Résultats estimés par SIFD pour la détection de plans

Tit Vidéo	OR	**V.P**		**V.N**		**F.P**	
	NB	NB	T%	NB	T%	NB	T%
15 vidéos PETS	84	62	73,8	22	18,03	16	13,11
112 vidéos propres	157	157	100	0	0	53	25,23

Tableau 3.2 – Résultats estimés par SIFD pour l'extraction des zones d'intérêts

toutes les vidéos et un taux de faux positifs **F.P** ne dépassant pas les 11 % parmi tous les changements de plan détectés.

Extraction des zones d'intérêts des vidéos

Comme pour la détection de plans, nous effectuons une comparaison des résultats de la procédure de l'extraction des zones d'intérêts du SIFD avec la vérité terrain. Nous présentons ces résultats dans le tableau 3.2. Ce dernier montre ceux obtenus pour les mêmes 15 vidéos de la collection PETS ainsi que ceux des 112 vidéos de notre collection. Dans ce tableau, nous retrouvons le nombre d'objets à détecter pour chaque groupe de vidéo selon la vérité terrain ainsi que celui des vrais positifs, des faux positifs et des vrais négatifs estimés par le SIFD. Le taux des vrais positifs est calculé par rapport au nombre d'objets à détecter. Par contre, le taux des Vrais Négatifs **V.N** et celui des Faux Positifs **F.P** sont calculés par rapport au nombre total d'objets détectés par SIFD.

Nous obtenons une bonne extraction des objets. À part $73,8$ % seulement trouvés dans le cas de la collection PETS, les objets des 112 vidéos propres sont tous biens segmentés. Cette diminution du taux de biens segmentés (vrai positif) pour la collection PETS est due surtout au nombre d'objets que chaque plan contient. Les vidéos de PETS contiennent plus d'un objet dans le même plan à l'opposé de notre collection où nous trouvons un seul objet. En analysant les résultats de la collection PETS, nous remarquons que lors de l'extraction des zones d'intérêts, les problèmes majeurs sont causés

par l'occultation par exemple dans les vidéos *OneShopOneWait2cor_P1*, *OneShopOneWait2cor_P2* et *OneShopOneWait2cor_P3* où trois personnes sont mals segmentées parce qu'elles marchent côte à côte dans la vidéo et il existe un contact entre elles. Ces trois personnes sont considérées comme un seul objet (voir figure 3.7).

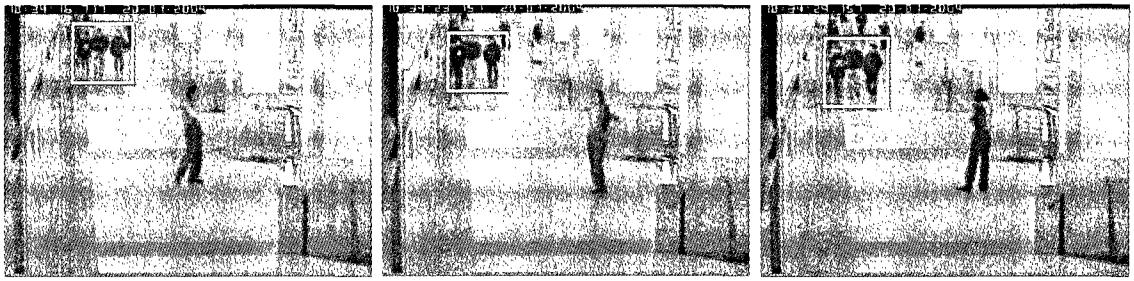


Figure 3.7 – *OneShopOneWait2cor_P1*- trois personnes sont détectés dans une seule zone

Nous remarquons aussi que les vrais négatifs sont causés par des changements de lumières. Dans les vidéos *Fight_OneManDown_P1*, *Fight_OneManDown_P2*, *Fight_RunAway2* et *LeftBag_AtChair_P1*, nous remarquons des objets détectés causés par le changement de lumière. Dans la vidéo *OneShopOneWait2cor_P1*, nous retrouvons ainsi un objet détecté causé par l'ombre 3.8.

Suivi des zones d'intérêts

Après l'extraction des zones d'intérêts des objets, nous effectuons leurs suivis. De même que pour les deux procédures précédentes, nous présentons les résultats obtenus pour

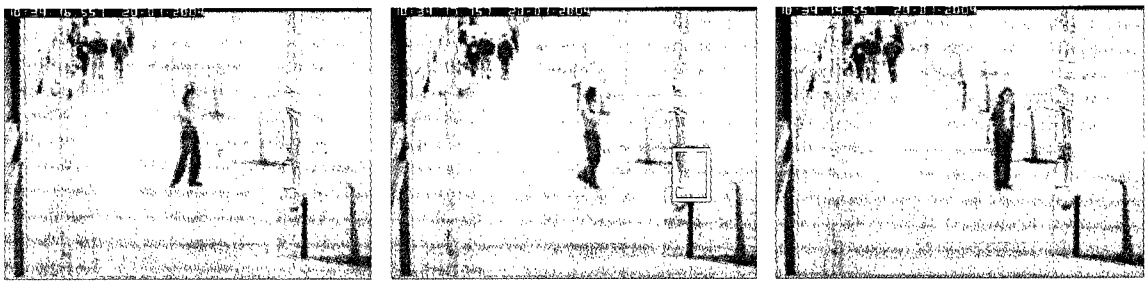


Figure 3.8 – *OneShopOneWait2cor_P1* - l'ombre de l'homme est détecté comme objet

Titre Vidéo	O.S	**B.S**		**N.S**		**M.S**	
	NB	NB	T%	NB	T%	NB	T%
15 vidéos PETS	84	63	75,00	0	0	21	25,00
112 vidéos propres	157	122	77,70	0	0	35	22,29

Tableau 3.3 – Résultats estimés par SIFD pour le suivi

les collections (PETS et vidéos propres) dans le tableau 3.3. Nous présentons dans ce dernier les performances de notre procédure de suivi par rapport à la vérité terrain et du fait le taux de réussite dans l'extraction de trajectoires. Dans les deux collections, nous observons que le nombre d'objets non suivis est égal à zéro. Tous les objets trouvés dans les vidéos sont généralement biens suivis **B.S** ou mals suivis **M.S**, d'où la trajectoire de l'objet est retrouvée complètement ou partiellement. Nous remarquons que les taux de biens suivis **B.S** et de mals suivis **M.S** pour les deux collections est de l'ordre de 75 % et 25 % respectivement. Après analyse des résultats du suivi, à part les problèmes reliés à la fausse détection des zones d'intérêts et la mauvaise détection de plan, d'autres problèmes tels que le problème du peu de points SIFT peuvent causer ce mauvais suivi. Par exemple, dans la vidéo *OneShopWait1front_P2* de la collection PETS, le suivi d'une femme à l'intérieur du magasin n'a pas pu être terminé, seulement une petite région de son corps se montrait et du fait le nombre de points d'intérêts dans cette zone détectée est insuffisant (figure 3.9).

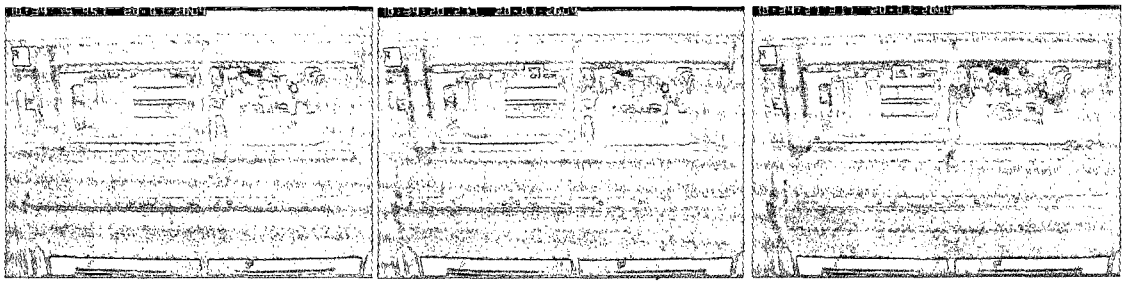


Figure 3.9 – *OneShopOneWait2cor_P1* - un nombre de points d'intérêts faible dans la zone d'intérêt pour effectuer le suivi

Tit Vidéo	A	B.R		M.R	
	NB	NB	T%	NB	T%
15 vidéos PETS	84	59	70,23	25	29,76
112 vidéos propres	157	120	76,43	37	23,56

Tableau 3.4 – Résultats estimés par SIFD pour la reconnaissance d'actions humaines

Reconnaissance d'actions humaines

Nous présentons dans le tableau 3.4, les résultats de la reconnaissance d'actions humaines pour les deux collections. Ce tableau contient le nombre d'actions reconnues par rapport à la vérité terrain ainsi que le nombre des actions mal reconnues ou inexistantes dans les actions d'apprentissage (Vrai Négatif). Ces résultats concernent les objets bien segmentés et dont la trajectoire est bien ou mal suivi.

Pour les deux collections, nous obtenons un taux de bien reconnu qui dépasse 70 % parmi les objets bien segmentés et dont la trajectoire est bien ou mal suivi, malgré que l'apprentissage s'effectue sur toutes les images de la vidéo et la reconnaissance à l'aide seulement des objets détectés et leurs trajectoires. Nous remarquons, aussi, que le taux

élevé d'actions mal reconnues pour la collection PETS est dû au manque d'actions à apprendre. Pour évaluer notre modèle de reconnaissance d'actions humaines appliqué sur des objets en mouvement, nous choisissons de présenter la matrice de confusion pour notre collection. Cette matrice permet de présenter le taux de bien classé pour chaque catégorie d'actions, ainsi que les catégories avec lesquelles une confusion se présente.

	Ouvrir porte	Se tenir debout	Courir	Marcher	S'asseoir
Ouvrir porte	**72,41%**	0%	13,79%	6,89%	6,89%
Se tenir debout	0%	**76,92%**	15,38%	7,69%	0%
Courir	0%	4,54%	**77,27%**	13,63%	4,54%
Marcher	0%	0%	11,76%	**82,35%**	5,8%
S'asseoir	0%	0%	14,28%	0%	**85,71%**

Tableau 3.5 – La matrice de confusion pour notre collection

La matrice de confusion pour nos 112 vidéos propres est présentée dans le tableau 3.5. Nous obtenons un taux moyen de bien classé de $78,93$ %, malgré que la reconnaissance se fait seulement sur les objets en mouvement et pas sur toute la vidéo comme dans le cas de l'apprentissage. Il est à remarquer que l'écart-type ne dépasse pas $4,62$, ce qui confirme la stabilité des résultats obtenus pour la reconnaissance d'actions humaines avec le modèle de classification MBRL étudié dans la section 2.3. Gloabalement, nous remarquons que plus les objets et leurs trajectoires sont bien définis, meilleurs sont les résultats de la reconnaissance d'actions humaines. Cette observation explique en partie que l'action « s'asseoir » est la mieux classée avec 85.71 % puisqu'elle est la mieux segmentée.

Nous comparons, dans le tableau 3.6, le taux d'actions bien classées par le SIFD à celui obtenu dans le travail [21]. Cette comparaison s'effectue sur les résultats obtenus par le SM2 sans la fusion des capteurs. Dans le travail de Chahid, le temps d'une action est manuellement définit à 3 secondes. Dans notre travail, cette contrainte n'existe pas et l'action est définie seulement par le mouvement de la personne.

Nous remarquons d'après le tableau 3.6 que SM2 obtient en général une meilleure recon-

	Ouvrir porte	Courir	S'asseoir	Se tenir debout	Marcher	**Moyenne**	**Écart-type**
SM2	69%	81,9%	91,7%	84,6%	85,3%	**82,5%**	**7,47**
SIFD	72,41%	77,27%	85,71%	76,92%	82,35%	**78,93%**	**4,62**

Tableau 3.6 – Comparaison entre les résultats obtenus par le SIFD et ceux obtenus par le système de surveillance par fusion de capteurs

Titre Vidéo	A	**B.R**		**M.R**	
	NB	NB	T%	NB	T%
112 vidéos propres	157	134	85,3	23	14,6

Tableau 3.7 – Résultats estimés par SIFD pour la reconnaissance d'actions humaines par apprentissage d'objets.

naissance d'actions. Dans le travail de Chahid, les vidéos d'apprentissage se retrouvent parmi les vidéos de tests sans aucun changement. Pour le SIFD, les éléments de tests ne sont plus des vidéos entières, mais seulement des objets en mouvement ce qui diffère des données d'apprentissage basées sur toute la vidéo en entrée. Nous remarquons aussi une amélioration pour la reconnaissance de l'action Ouvrir porte. Cela est dû à l'extraction des zones d'intérêts qui distingue bien la zone de la porte et le mouvement de l'ouverture de la porte.

Nous testons l'impact de l'apprentissage à partir d'objets au lieu de l'apprentissage à partir de toute la vidéo. Nous avons choisi d'apprendre un nombre total de 39 objets, dont 6 pour l'action « marcher », 6 pour « courir », 12 pour « ouvrir une porte », 6 pour « se tenir debout » et 9 objets pour l'action « s'asseoir » sur une chaise. Nous testons SIFD sur les 112 vidéos de notre base de données. Dans le tableau 3.7, nous remarquons une amélioration de la reconnaissance de l'ordre de $6,37\%$.

D'après les résultats obtenus par les quatre procédures ci-dessus, nous pouvons conclure, qu'avec $85,3$ % d'actions reconnues pour notre collection et 70 % pour la collection PETS, le SIFD n'a besoin d'aucune intervention humaine dans le processus de reconnaissance

à l'encontre du système pour la surveillance par fusion de capteurs qui a besoin d'intervention humaine pour définir les segments de vidéos pour chaque action. SIFD éprouve des problèmes surtout lors de l'extraction des zones d'intérêts causés entre autres par les changements de luminosité et les problèmes d'occultations.

Exemples de dictionnaire vidéos

Nous présentons, ci-dessous, les dictionnaires obtenus pour deux vidéos de domaines différents. La première vidéo est un extrait du film *OneStopNoEnter1cor* d'une durée de 28 secondes en raison de 25 images/seconde 3.10. La deuxième vidéo est un extrait de la série *Friends* d'une durée de 600 secondes et de l'ordre de 25 images/seconde.

Le tableau 3.8 représente le dictionnaire de la vérité terrain de la vidéo *OneStopNoEnter1cor* et le tableau 3.9 représente celui estimé par le SIFD. Pour la procédure de l'extraction des zones d'intérêts, le SIFD détecte 3 objets illustrés dans le tableau 3.9 parmi les 5 objets présentés dans le dictionnaire de la vérité terrain dans le tableau 3.8. Nous remarquons la disparition des deux derniers objets O4 et O5 du dictionnaire estimé. Cela est dû principalement au mouvement lent de ces objets. La détection du contour spatio-temporel ne distingue pas le mouvement puisque la zone de changement est nulle. Pour la reconnaissance d'actions, nous considérons que l'action « se tenir debout » trouvée par notre système coïncide avec l'action arrêt. Toutes actions détectées ont été suivies.

Le problème montré par cette expérimentation est la mauvaise détection d'objets qui se déplacent à une faible vitesse. Notre SIFD s'exécute en 3300 secondes pour effectuer l'extraction des zones d'intérêts et le suivi et en 18 secondes pour effectuer la reconnaissance d'actions pour la vidéo *OneStopNoEnter1cor*. Ce temps d'exécution est dû principalement à l'extraction des caractéristiques, surtout les points d'intérêts SIFT pour effectuer le suivi.

Figure 3.10 – Exemple d'images de la vidéo *OneStopNoEnter1cor*

NM=*OneStopNoEnter1cor.mpg*, NB=700, IW=360, IH=240

[[**Pl** :0, **IR**=  , **ID**=0 ,**IF**=700]

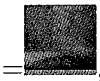
[**O** 1, **IO**=  , **OD**=0, **OF**=52, **Mv**=marcher]

[**O** 1, **IO**=  , **OD**=53, **OF**=95, **Mv**=arret]

[**O** 1, **IO**=  , **OD**=96, **OF**=393, **Mv**=marcher]

[**O** 2, **IO**=  , **OD**=0, **OF**=700, **Mv**=marcher]

[**O** 3, **IO**=  , **OD**=97, **OF**=700, **Mv**=marcher]

[**O** 4, **IO**=  , **OD**=340, **OF**=400, **Mv**=bouger tete]

[**O** 5, **IO**=  , **OD**=343, **OF**=400, **Mv**=bouger tete]]

Tableau 3.8 – Dictionnaire de la vérité terrain de la vidéo *OneStopNoEnter1cor* .

NM=*OneStopNoEnter1cor.mpg*, NB=700, IW=360, IH=240

[[**Pl** :0, **IR**=  , **ID**=0 ,**IF**=700]

[**O** 1, **IO**=  , **OD**=0, **OF**=52, **Mv**=marcher]

[**O** 1, **IO**=  , **OD**=53, **OF**=95, **Mv**=arret]

[**O** 1, **IO**=  , **OD**=96, **OF**=393, **Mv**=marcher]

[**O** 2, **IO**=  , **OD**=0, **OF**=700, **Mv**=marcher]

[**O** 3, **IO**=  , **OD**=97, **OF**=700, **Mv**=marcher]

Tableau 3.9 – dictionnaire estimé automatiquement de la vidéo *OneStopNoEnter1cor*.

Figure 3.11 – Une épisode de la série *Friends* de 600 secondes avec 25 images/seconde

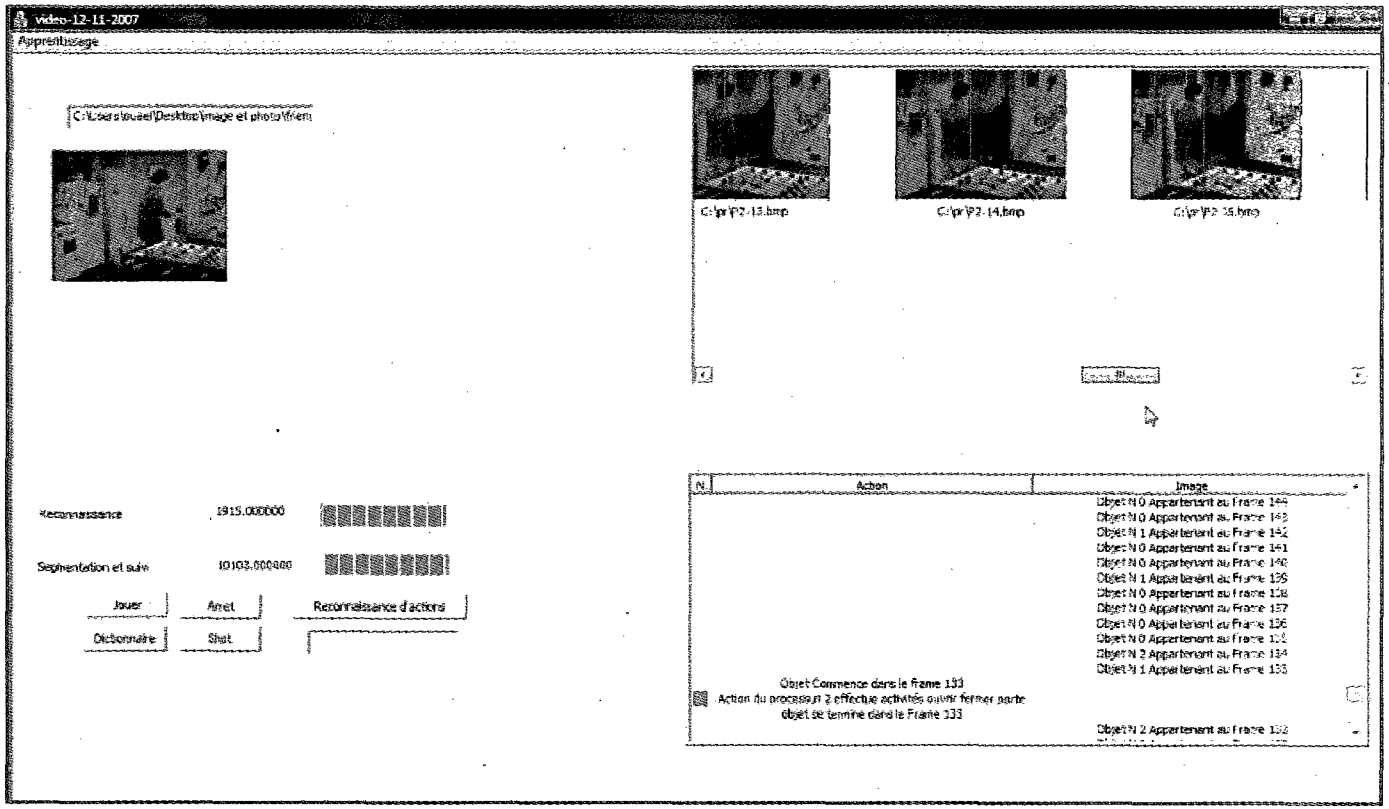


Figure 3.12 – Extrait du programme et des résultats de Friends

Nous présentons en ce qui suit les résultats de la vidéo d'un épisode de la série *Friends* tel l'extrait dans la figure 3.12 dont un résumé est illustré dans la figure 3.11. Dans le tableau 3.10, nous illustrons le nombre de plans détectés par le système SIFD et la vérité terrain construite manuellement. Nous présentons les scores dans le tableau 3.10 et nous remarquons que tous les plans (Vrais Positifs) ont été détectés avec un pourcentage de faux positifs de l'ordre de $9,5\%$.

Nous montrons dans le tableau 3.11 les résultats obtenus à partir de notre extraction

Tit Vidéo	**OR**	**V.P**	**V.N**	**F.P**
Friends	314	314	0	30

Tableau 3.10 – Résultats estimés par SIFD pour la détection de plans

des zones d'intérêts. Nous obtenons un taux de $84,05$ de zones d'intérêts retrouvées et un taux de 15% de zones non détectées (Vrais Négatifs) ou qui ont été détectées sans qu'elle existe (Faux Positifs). Cela revient généralement au problème d'occultation surtout dans les plans qui comportent plusieurs zones. Par exemple, des zones sont fusionnées en une seule zone telle présentée dans la figure 3.13. Nous remarquons que la femme et l'enfant sont fusionnés dans une seule zone d'intérêts de même pour l'homme et la serveuse derrière lui.

Nous vérifions en ce qui suit les résultats du suivi effectué par le SIFD avec ceux de la vérité terrain. Nous illustrons nos résultats dans le tableau 3.12. Parmi les zones d'intérêts bien détectés ou vrais positifs, nous cherchons celles où le suivi était bien effectué (**B.S** : Biens Suivis) par rapport à la vérité terrain. Nous obtenons un taux de **B.S** de $77,24\%$ et de $22,75$ de mals suivis **M.S** pour un total de 100%. Pour toutes les zones bien segmentées,

Figure 3.13 – Extraction des objets dans une scène de la série : O1 fille qui entre au magasin, O2 l'homme à droite debout entrain de parler, O3 l'homme a droite assis sur la table et qui bouge la tête.

Tit Vidéo	**OR**	**V.P**		**V.N**		**F.P**	
	NB	NB	T%	NB	T%	NB	T%
Friends	345	290	84,05	50	14,49	60	15,00

Tableau 3.11 – Résultats estimés par SIFD pour l'extraction des zones d'intérêts

le suivi est effectué. Les zones **M.S** sont des zones où des ruptures de suivi sont causées par le peu de points d'intérêts retrouvés surtout pour les zones qui contiennent un petit mouvement.

Pour la reconnaissance d'actions humaines, nous utilisons les mêmes données d'apprentissages présentés ci-dessus (notre collection). Pour fabriquer la vérité terrain, nous catégorisons manuellement les actions de cet épisode de la série *Friends* selon 6 actions « marcher », « courir », « s'asseoir », se tenir debout, ouvrir porte et autres actions

Titre Vidéo	O.S	**B.S**		**N.S**		**M.S**	
	NB	NB	T%	NB	T%	NB	T%
Friends	290	224	77,24	0	0	64	22,75

Tableau 3.12 – Résultats estimés par SIFD pour le suivi

or celles connues pour l'apprentissage. Le tableau 3.13 illustre une comparaison entre la reconnaissance d'actions humaines par notre SIFD avec la vérité terrain. Nous obtenons 74, 69% des actions sont reconnues par SIFD parmi toutes les actions de la vérité terrain.

	Ouvrir porte	Courir	S'asseoir	Se tenir debout	Marcher	Autre
Verité Terrain	6	0	43	99	14	128
SIFD	4	0	36	71	10	128

Tableau 3.13 – Comparaison entre les résultats obtenus par le SIFD et la vérité terrain

Temps d'exécution du SIFD

Dans le tableau 3.14 nous présentons les moyennes des temps d'exécution en secondes pour les quatre procédures de notre SIFD pour les trois collections étudiées. Les temps d'exécution (en secondes) de l'extraction des zones d'intérêts et du suivi sont présentés ensemble puisque les deux s'exécutent ensemble dans notre application. Comme présenté dans la section **collection**, nous avons 15 vidéos de PETS avec chacune 400 images de taille 360 × 240 pixels. Notre collection contient, aussi, 112 vidéos avec 200 images de 360 × 250 pixels. L'épisode de *Friends* contient à elle-même 15000 images de 160 × 120. D'après le tableau 3.14 nous remarquons que l'extraction des zones d'intérêts de l'épisode de Friends a un temps d'extraction des zones d'intérêts et de suivi plus rapide que les autres vidéos en temps de traitement (secondes) par image. Nous obtenons une moyenne du temps de traitement pour l'extraction des zones d'intérêts et le suivi (segementation) d'environ 3 secondes par image tandis que pour les 112 vidéos propres, la moyenne est de 11 secondes/image et 17 secondes pour la collection PETS. Cela s'explique entre autres par la taille de l'image utilisée dans *Friends* qui est plus petite que les deux autres. Nous pouvons aussi expliquer le temps plus long pour les vidéos PETS par le nombre d'objets contenu qui est plus grand que celui des 112 vidéos propres. Ces derniers contiennent généralement 1 à 2 actions maximums avec une moyenne de 5 secondes pour

la reconnaissance d'actions. Pour les vidéos PETS, elles contiennent en moyenne entre 3 et 4 objets par vidéo avec une moyenne de temps d'exécution par vidéo de 15 secondes pour la reconnaissance. Nous obtenons en moyenne entre 3 et 7 secondes de temps de reconnaissance pour chaque objet. Le temps d'exécution assez lent pour l'extraction des zones d'intérêts revient généralement à l'extraction de caractéristiques entre autres le gradient spatio-temporel et les points SIFT.

Titre Vidéo	Plans		Seg, Suivi		Actions		Temps d'exécution	
	Moy	Ecart	Moy	Ecart	Moy	Ecart	Moy	Ecart
15 vidéos PETS	56	10	6800	250	15	7	8000	520
112 vidéos propres	23	5	2200	171	5	2	2350	256
Friends	402	0	45000	0	150	0	55000	0

Tableau 3.14 – Moyenne et écart-type du temps d'exécution en secondes du SIFD par vidéo

3.5 Conclusion

Nous avons illustré dans ce chapitre, la production du dictionnaire des événements de la vidéo. Le SIFD consiste à détecter automatiquement, sans l'intervention de l'être humain, les événements d'une vidéo.

Nous procédons tout d'abord à une détection de coupure de plan pour ensuite segmenter chaque séquence d'images appartenant au même plan. Cette extraction des zones d'intérêts se base sur la détection du contour spatio-temporel. Puis, nous effectuons un suivi des objets pour former leurs trajectoires. Nous reconnaissons pour chaque objet en mouvement trouvé, l'action qu'il définit.

Nous avons testé notre dictionnaire sur des vidéos, celles provenant d'un système de

vidéosurveillance (15 vidéos PETS et 112 vidéos propres) et celle de la série *Friends*. Nous avons obtenu respectivement une précision qui dépasse les 97%, 100% et 85% pour la détection de plan, les 73%, 100% et 84% pour l'extraction des zones d'intérêts, les 100%, 100% et 100% pour le suivi bien et mal effectué et les 70%, 76% et 74% pour l'extraction des zones d'intérêts afin de construire le dictionnaire qui décrit les événements des vidéos. Il reste que le dictionnaire dépend de la collection des actions d'apprentissage.

Pour améliorer les performances d'un tel système, il faut tout d'abord développer un meilleur apprentissage avec plus d'actions à reconnaitre. Il est intéressant aussi d'améliorer l'extraction des objets et surtout le contour spatio-temporel, en ajoutant une information sur la forme de l'objet a priori. Une nette amélioration du temps d'exécution pourra aussi intéressante avèc un gradient avec les masques de Sobel (à la place de la gaussienne).

Comme perspective à long terme, nous proposons l'élaboration d'un système plus générale. Ce système intègrera, par exemple, l'audio et la mise en relation d'événements pour améliorer la précision du dictionnaire.

CONCLUSION ET PERSPECTIVES

Puisque l'utilisation de la vidéo devient aujourd'hui incontournable dans de nombreux domaines. Parfaire son interprétation est d'un grand intérêt, surtout quand il s'agit de l'analyse du comportement humain, un secteur de recherche en plein essor dans la communauté de la vision par ordinateur. Dans le but d'améliorer l'interprétation des vidéos et la reconnaissance d'actions humaines, nous avons ciblé la fabrication d'un dictionnaire. L'approche proposée se base sur quatre étapes qui permettent la formation d'un dictionnaire garantissant une description détaillée des actions dans la vidéo. Ce sont la coupure de plans, l'extraction des zones d'intérêts, le suivi d'objets et la reconnaissance d'actions humaines. Dans ce travail, l'approche de reconnaissance d'actions humaines que nous avons développée est comparée à d'autres approches. Notre contribution se situe au niveau de l'extraction de caractéristiques, de la formation du dictionnaire de l'implantation et de la validation. Pour l'extraction de caractéristiques, nous avons proposé une nouvelle caractéristique appelée « CSST », qui combine deux caractéristiques les « PIST » et les « CST ». Les résultats obtenus ont donné le meilleur taux de reconnaissance et le meilleur écart-type de ce taux par rapport à des travaux existants dans la littérature. Ce travail a été réalisé conjointement avec Omar Chahid du Centre Moivre.

Pendant la formation du dictionnaire, une combinaison de plusieurs techniques dans le but d'interpréter la vidéo a été fournie et une trace des objets et de leurs trajectoires a été

gardée afin de permettre une meilleure extraction des zones d'intérêts et suivi d'objets. Les exemples de dictionnaire estimés montrent des résultats plus au moins proches de la vérité terrain et fournissent une bonne idée sur les événements dans une vidéo. Notre système a l'avantage d'être automatique, d'être non spécifique à un seul domaine. Notre approche ouvre de nombreuses perspectives d'amélioration et d'extension du système. Il sera intéressant d'inclure un modèle de mise à jour dans le modèle de reconnaissance d'actions humaines, dans le but d'ajouter de nouvelles actions au fur et à mesure que le système fabrique des dictionnaires. Ainsi, le manque de données pour l'apprentissage peut être résolu. En outre, inclure l'étude d'actions non humaines, comme le mouvement de voiture serait également intéressant. Avec cet ajout, un système de vidéosurveillance pourrait avoir un compte-rendu, en forme de dictionnaire, du mouvement de la circulation.

Bibliographie

[1] Gestion Électronique des documents. Wikipédia l'encyclopédie libre. URL http ://fr.wikipedia.org/wiki/Gestion_électronique_des_documents.

[2] Indigovision's analytics algorithms. IndigoVision, Complete IP Video Security Solutions. URL http ://www.indigovision.com.

[3] MARS Exploration Rover Mission . NASA, Jet Propulsion Laboratory California Institute Of Technology. URL http ://marsrovers.nasa.gov/home/index.html.

[4] Youtube statistics. Digital Ethnography at KSU Project Wiki. URL http ://ksudigg.wetpaint.com/page/YouTube+Statistics.

[5] Analyse sur l'enquête internationale de l'institut des statistiques de l'unesco sur les statistiques de films de long métrage. ISU, 2007.

[6] Pets benchmark data, 2007. URL http ://www.cvg.rdg.ac.uk/PETS2007/data.html.

[7] Statistiques sur l'industrie du film et de la production télévisuelle indépendante, édition 2008. Institut de la Statistique Québec, 2008.

[8] E.H. ADELSON et J.Y.A. WANG. Representing moving images with layers. *IEEE Transactions on Image Processing*, pages 625–638, 1994.

[9] J.K. AGGARWAL et Q. CAI. Human motion analysis : A review. *Computer Vision and Image Understanding*, pages 428–440, 1999.

[10] P. AIGRAIN et P. JOLY. The automatic real-time analysis of film editing and transition effects and its applications. *Computers And Graphics*, pages 93–103, 1994.

[11] M.S. ALLILI et D. ZIOU. Active contours for video object tracking using region, boundary and shape information. *Signal, Image and Video Processing*, pages 101–117, 2007.

[12] N. BABAGUCHI et R. JAIN. Event detection from continuous media. pages 1209–1212, 1998.

[13] J. BEN-ARIE, P. Zhiqian Wang PANDIT et S. RAJARAM. Human activity recognition using multidimensional indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1091–1104, 2002.

[14] Y. BENEZETH, P.M. JODOIN, B. EMILE, H. LAURENT et C. ROSENBERGER. Review and evaluation of commonly-implemented background subtraction algorithms. pages 1–4, 2008.

[15] A.F. BOBICK et J.W. DAVIS. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 257–267, 2001.

[16] G. BOCCIGNONE, A. CHIANESE, V. MOSCATO et A. PICARIELLO. Foveated shot detection for video segmentation. *IEEE Transaction on Circuits and Systems for Video Technology*, pages 365–377, 2005.

[17] G.D. BORSHUKOV, G. BOZDAGI, Y. ALTUNBASAK et A.M. TEKALP. Motion segmentation by multi-stage affine classification. *IEEE Transactions on Image Processing*, pages 1591–1594, 1997.

[18] P. BOUTHEMY et E. FRANÇOIS. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *International Journal of Computer Vision*, pages 157–182, 1993.

[19] M. BRAND et V. KETTNAKER. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 844–851, 2000.

[20] C. BREGLER. Learning and recognizing human dynamics in video sequences. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 568–574, 1997.

[21] O. CHAHID. Détection d'actions humaines interdites par fusion de capteurs. Mémoire de Maîtrise, Université de Sherbrooke, 2009.

[22] O. CHOMAT et J.L. CROWLEY. Probabilistic recognition of activity using local appearance. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 104–109, 1999.

[23] R. COLLINS, A. LIPTON et T. KANADE. A system for video surveillance and monitoring. *American Nuclear Society 8th Internal Topical Meeting on Robotics and Remote Systems*, 1999.

[24] D. CREMERS, M. ROUSSON et R. DERICHE. A review of statistical approaches to level set segmentation : Integrating color, texture, motion and shape. *International Journal of Computer Vision*, pages 195–215, 2007.

[25] N. CRISTIANINI et J. SHAWE-TAYLOR. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[26] J.E. CUTTING et L.T. KOZLOWSKI. Recognizing friends by their walk : Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, pages 353–356, 1977.

[27] J. DEMONGEOT, G. VIRONE, F. DUCHENE, G. BENCHETRIT, T. HERVE, N. NOURY et V. RIALLE. Multi-sensors acquisition, data fusion, knowledge mining and triggering in health smart homes for elderly people. *Comptes Rendus Biologies*, pages 673–682, 2002.

[28] D.G. LOWE. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004.

[29] P. DOLLAR, V. RABAUD, G. COTTRELL et S. BELONGIE. Behavior recognition via sparse spatio-temporal features. *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[30] N. FRIEDMAN, K. MURPHY et S. RUSSELL. Learning the structure of dynamic probabilistic networks. *Conference on Uncertainty in Artificial Intelligence*, pages 139–147. Morgan Kaufmann, 1998.

[31] V. GOUAILLIER et A.E. FLEURANT. La vidéosurveillance intelligente : promesses et défis. Rapport technique, TechnoPole Défense and Security, CRIM, 2009.

[32] Y. GUO, G. XU et S. TSUJI. Understanding human motion patterns. *IEEE International Conference on Pattern Recognition*, pages 325–329, 1994.

[33] A. HAKEEM et M. SHAH. Ontology and taxonomy collaborated framework for meeting classification. *IEEE International Conference on Pattern Recognition*, pages 219–222, Washington, DC, USA, 2004. IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[34] A. HAKEEM et M. SHAH. Learning, detection and representation of multi-agent events in videos. *American Association of Artificial Intelligence*, pages 586–605, 2007.

[35] I. HARITAOGLU, D. HARWOOD et L. DAVIS. W4 : Who, when, where, what : A real time system for detecting and tracking people. *IEEE International Conference on Automatic Face and Gesture recognition*, pages 222–227, 1998.

[36] I. HARITAOGLU, D. HARWOOD et L.S. DAVIS. Ghost : A human body part labeling system using silhouettes. *IEEE International Conference on Pattern Recognition*, pages 77–82, 1998.

[37] C. HARRIS et M. STEPHENS. A combined corner and edge detection. *The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[38] M. HARVILLE et Dalong L.. Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 398–405, 2004.

[39] C.T. HSU et Y.C. TSAN. Mosaics of video sequences with moving objects. *IEEE International Conference on Image Processing*, pages 387–390, 2001.

[40] Y. HUANG. An iterative approach for segmenting video objects under occlusion. *IEEE International Symposium on Intelligent Information Technology Applications*, pages 442–446, 2008.

[41] J. HUART. *Extraction et analyse d'objets-clés pour la structuration d'images et de vidéos.* Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG, 2007.

[42] Y.A. IVANOV et A.F. BOBICK. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 852–872, 2000.

[43] G. JOHANSSON. Visual motion perception. *Scientific American*, pages 76–88, 1975.

[44] Y. KE, R. SUKTHANKAR et M. HEBERT. Efficient visual event detection using volumetric features. *IEEE International Conference on Computer Vision*, pages 166–173, 2005.

[45] W. KIENZLE, B. SCHOLKOPF, F.A. WICHMANN et M.O. FRANZ. How to find interesting locations in video : A spatiotemporal interest point detector learned from human eye movements. *Symposium of the German Association for Pattern Recognition*, pages 405–414, 2007.

[46] M. KOHLE, D. MERKL et J. KASTNER. Clinical gait analysis by neural networks : Issues and experiences. *IEEE Symposium on Computer-Based Medical Systems*, pages 138–143, 1997.

[47] D. KOLLER, H. HEINZE et H.H. NAGEL. Algorithmic characterization of vehicle trajectories from image sequences by motion verbs. *Computer Vision and Pattern Recognition*, pages 90–95, 1991.

[48] N. KRAHNSTOVER, M. YEASIN et R. SHARMA. Towards a unified framework for tracking and analysis of human motion. *IEEE Workshop on Detection and Recognition of Events in Video*, pages 47–54, 2001.

[49] I. LAPTEV et T. LINDEBERG. Space-time interest points. *IEEE International Conference on Computer Vision*, pages 432–439, 2003.

[50] I. LAPTEV, M. MARSZALEK, C. SCHMID et B. ROZENFELD. Learning realistic human actions from movies. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[51] S. LAWRENCE, D. ZIOU, M.-F. AUCLAIR-FORTIER et S. WANG. Motion insensitive detection of cuts and gradual transitions in digital videos. *Pattern Recognition and Image Analysis*, pages 109–119, 2004.

[52] V. LEPETIT, A. SHAHROKNI et P. FUA. Robust data association for online application. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 281–288, 2003.

[53] F. LERASLE, G. RIVES et M. DHOME. Tracking of human limbs by multiocular vision. *Computer Vision Image Understanding*, pages 229–246, 1999.

[54] M.K. LEUNG et Y.-H. YANG. First sight : A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 359–377, 1995.

[55] D. LÉVESQUE et F. DESCHÊNES. Sparse scene structure recovery from atmospheric degradation. *IEEE International Conference on Pattern Recognition*, pages 84–87, 2004.

[56] D. LÉVESQUE et F. DESCHÊNES. Novel depth cues from light scattering. *Image and Vision Computing*, pages 19–36, 2009.

[57] P. MAES, T. DARRELL, B. BLUMBERG et A. PENTLAND. The alive system : Wireless, full-body interaction with autonomous agents. *Multimedia Systems*, pages 105–112, 1997.

[58] N. MAILLOT, M. THONNAT et A. BOUCHER. Towards ontology-based cognitive vision. *Machine Vision and Applications*, pages 33–40, 2004.

[59] O. MASOUD et N. PAPANIKOLOPOULOS. A method for human action recognition. *Image and Vision Computing*, pages 729–743, 2003.

[60] H. MENG, N. PEARS et C. BAILEY. A human action recognition system for embedded computer vision application. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[61] D. MEYER, J. DENZLER et H. NIEMANN. Model based extraction of articulated objects in image sequences for gait analysis. *IEEE International Conference on Image Processing*, pages 78–81, 1997.

[62] T.B. MOESLUND et E. GRANUM. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, pages 231–268, 2001.

[63] A. NAGASAKA et Y. TANAKA. Automatic video indexing and full-video search for object appearances. *Conference on Visual Database Systems II*, pages 113–127, Amsterdam, The Netherlands, 1992. North-Holland Publishing Co.

[64] J. NAM et A.H. TEWFIK. Detection of gradual transitions in video sequences using b-spline interpolation. *IEEE Transactions on Multimedia*, pages 667–679, 2005.

[65] M.R. NAPHADE, R. MEHROTRA, A.M. FERMAN, J. WARNICK, T.S. HUANG et A.M. TEKALP. A high-performance shot boundary detection algorithm using multiple cues. *IEEE International Conference on Image Processing*, pages 884–887, 1998.

[66] J. NIEBLES, H. WANG et L. FEI-FEI. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, pages 299–318, 2008.

[67] E. ONG et S. GONG. Tracking hybrid 2d-3d human models from multiple views. *IEEE International Workshop on Modelling People*, pages 11–18, 1999.

[68] M. PETKOVIC, W. JONKER et Z. ZIVKOVIC. Recognizing strokes in tennis videos using hidden markov models. *International Conference on Visualization, Imaging and Image Processing*, pages 512–516, 2001.

[69] R. POLANA et R. NELSON. Detecting activities. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2–7, 1993.

[70] G. QIAN, S. SURAL, Y. GU et S. PRAMANIK. Similarity between euclidean and cosine angle distance for nearest neighbor queries. *ACM Symposium on Applied Computing*, pages 1232–1237, 2004.

[71] B. Colin R. KSANTINI, D. Ziou et F. DUBEAU. Weighted pseudometric discriminatory power improvement using a bayesian logistic regression model based on a variational method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 253–266, 2007.

[72] B. Colin R. KSANTINI, D. Ziou et F. DUBEAU. A bayesian kernel logistic discriminant model : An improvement to the kernel fisher's discriminant. *National Conference on American Association of Artificial Intelligence*, pages 1464–1465, 2008.

[73] C. RAO, A. YILMAZ et M. SHAH. Iview-invariant representation and recognition of actions, 2002.

[74] T. REMI et M. BERNARD. Probabilistic matching algorithm for keypoint based object tracking using a delaunay triangulation. *IEEE International Workshop on Image Analysis for Multimedia Interactive Services*, pages 17–17, 2007.

[75] Y. RUI et P. ANANDAN. Segmenting visual actions based on spatio-temporal motion patterns. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 111–118, 2000.

[76] H. SAWHNEY et S. AYER. Compact representations of videos through dominant and multiple motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 814–830, 1996.

[77] C. SCHÜLDT, I. LAPTEV et B. CAPUTO. Recognizing human actions : A local svm approach. *IEEE International Conference on Pattern Recognition*, pages 32–36, 2004.

[78] C. SCHMID et R. MOHR. Image retrieval using local characterization. *IEEE International Conference on Image Processing*, pages 781–784, 1996.

[79] I.K. SETHI et R. JAIN. Finding trajectories of feature points in a monocular image sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 56–73, 1987.

[80] J.A. SETHIAN. *Level Set Methods and Fast Marching Methods : Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science.* Cambridge University Press, 1999.

[81] T.E. STARNER, J. WEAVER et A.P. PENTLAND. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1371–1375, 1998.

[82] N. VASCONCELOS et A. LIPPMAN. A spatiotemporal motion model for video summarization. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 361–366, 1998.

[83] C.J. VEENMAN, M.J.T. REINDERS et E. BACKER. Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 54–72, 2001.

[84] L. WANG, W. HU et T. TAN. Recent developments in human motion analysis. *Pattern recognition*, pages 585–601, 2003.

[85] S.F. WONG et R. CIPOLLA. Extracting spatiotemporal interest points using global information. *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[86] J. YANG et A. WAIBEL. A real-time face tracker. *IEEE Workshop on Applications of Computer Vision*, pages 142–147, 1996.

[87] C. YEO, P. AHAMMAD, K. RAMCH et S.S. SASTRY. Compressed domain real-time action recognition. *IEEE Workshop on Multimedia Signal Processing*, pages 33–36, 2006.

[88] A. YILMAZ, O. JAVED et M. SHAH. Object tracking : A survey. *ACM Computing Surveys*, pages 13–17, 2006.

[89] R. ZABIH, J. MILLER et K. MAI. A feature-based algorithm for detecting and classifying production effects. pages 119–128, 1999.

[90] L. ZELNIK-MANOR et M. IRANI. Event-based analysis of video. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 123–123, 2001.

[91] L. ZELNIK-MANOR et M. IRANI. Statistical analysis of dynamic actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1530–1535, 2006.

[92] H.J. ZHANG, A. KANKANHALLI et S.W. SMOLIAR. Automatic partitioning of full-motion video. *Multimedia Systems*, pages 10–28, 1993.

[93] H. ZHONG, J. SHI et M. VISONTAI. Detecting unusual activity in video. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2004.